

Improved Approximation of Storage-Rate Tradeoff for Caching with Multiple Demands

Avik Sengupta, *Student Member, IEEE*, Ravi Tandon, *Member, IEEE*

Abstract

Caching at the network edge has emerged as a viable solution for alleviating the severe capacity crunch in modern content centric wireless networks by leveraging network load-balancing in the form of localized content storage and delivery. In this work, we consider a cache-aided network where the cache storage phase is assisted by a central server and users can demand multiple files at each transmission interval. To service these demands, we consider two delivery models - (1) centralized content delivery where user demands at each transmission interval are serviced by the central server via multicast transmissions; and (2) device-to-device (D2D) assisted distributed delivery where users multicast to each other in order to service file demands. For such cache-aided networks, we present new results on the fundamental cache storage vs. transmission rate tradeoff. Specifically, we develop a new technique for characterizing information theoretic lower bounds on the storage-rate tradeoff and show that the new lower bounds are strictly tighter than cut-set bounds from literature. Furthermore, using the new lower bounds, we establish the optimal storage-rate tradeoff to within a constant multiplicative gap. We show that, for multiple demands per user, achievable schemes based on repetition of schemes for single demands are order-optimal under both delivery models.

I. INTRODUCTION

The dynamics of traffic over wireless networks has undergone a paradigm shift to become increasingly content centric with high volume multimedia content (e.g., video) distribution holding precedence. Therefore, efficient utilization of network resources is imperative in such networks for improving capacity. With the proliferation of cheap storage at the network edge (e.g. at user devices and small cell base stations), caching has emerged as an important tool for facilitating efficient load balancing and maximal resource utilization for future 5G wireless networks [1]. Parts of popular files are pre-stored at the edge caches such that at times of high network load, the local content can be leveraged to reduce the over-the-air transmission rates. Caching and complimentary file delivery in wireless networks has been the subject of a wealth of recent research as evidenced by the results in [2]–[17]. Caching generally works in two phases - (a) the *cache storage phase* where parts of popular content is placed in users' cache memories by a central server e.g., an LTE eNodeB in modern cellular networks and (b) the *file delivery phase*, where requested content is delivered by exploiting local cache storage. Cache placement happens over a much larger time-scale than the file request and delivery phase or a *transmission interval*, and needs to be agnostic to user demands. The fundamental tradeoff in such cache-aided systems is between the cache storage and the delivery rate.

Recently, Maddah-Ali and Niesen [2]–[5] showed that by jointly designing the storage and delivery phases, order-wise improvement in the delivery rate can be achieved for any given size of cache storage for the case when users demand only one file at every transmission interval. The proposed schemes extract a *global caching gain*, in addition to the traditional *local caching gain*, by distributing common content across users' caches and subsequently designing *centralized coded multicast transmissions* which leverage this shared content to reduce delivery rates. The authors used cut-set based arguments to derive an information theoretic lower bound on the optimal storage-rate tradeoff and characterized it to within

A. Sengupta is with Wireless@VT and The Hume Center for National Security and Technology, Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060 USA. Email: aviksg@vt.edu. R. Tandon is with the Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721 USA. Email: tandonr@email.arizona.edu.

This work was presented in part at the IEEE International Symposium of Information Theory (ISIT), Hong Kong, June 2015 and at the Information Theory and Applications Workshop (ITA), UCSD, Feb 2015.

a constant multiplicative factor of 12 for worst-case user demands under uniform file popularity. An new lower bound as well as an improved characterization of the optimal storage-rate tradeoff to within a factor of 8 was presented in our previous work in [18]. The case when users demand multiple files at each transmission interval was initially studied in [19], [20] for the case of worst-case as well as random user demands. The authors in [19] also proposed a cut-set lower bound for this setting.

In contrast to the centralized delivery model, a distributed device-to-device (D2D) assisted delivery model was studied in [11] whereby the delivery phase was relegated to the users instead of a centralized server in order to further reduce backhaul load. The main difference between the centralized content delivery studied in [2], [3] and the D2D-assisted delivery studied in [11] is the *distributed nature of multicast transmissions*. In the centralized delivery model of Maddah-Ali and Niesen, the multicast can be any arbitrary function of all the files in the library. Instead, for D2D-assisted delivery, the outgoing multicast from each user *can only depend* on the local cache content of that device. In [11], Ji et.al. presented new storage/delivery mechanisms for D2D-assisted delivery for the case when each user demands a single file at every transmission interval. The results in [11] show that even for D2D-assisted delivery, when the devices can use inter-device coded multicast transmissions to satisfy the demands of other users, order-wise improvements in terms of delivery rate can be achieved as compared to uncoded delivery. The authors also presented a cut-set based lower bound on the storage-rate tradeoff. In our prior work in [21], we improved on the cut-set bound and showed that the achievable scheme in [11] is within a constant multiplicative factor of 8 from the optimal by leveraging the new bounds. However, the general case when each user can demand *multiple files* at each transmission interval with D2D-assisted delivery has not been considered in literature.

Main Contributions: The main contributions of the paper are summarized as follows.

- We develop a *new technique* for characterizing information theoretic lower bounds on the storage-rate trade-off for cache-aided systems under centralized and D2D-assisted content delivery for the general case when users can demand multiple files at each transmission interval.
- The new lower bounds are shown to be generally tighter than the cut-set bounds in [2, Theorem 2] and [19, Theorem 2] for all values of problem parameters. The proposed technique also yields the first known converse for the case of D2D-assisted delivery when each user demands multiple files at each transmission interval.
- Using the new lower bounds we show that repetitive use of the achievable schemes for single file demands per user, proposed in [2, Theorem 1] and [11, Theorem 1], is order-optimal for the case of multiple demands and can characterize the optimal storage-rate tradeoff to within a constant multiplicative factor of 11 for centralized delivery and 10 for D2D-assisted delivery, improving upon known results.

Notation: For any two integers a, b with $a \leq b$, we define $[a : b] \triangleq \{a, a + 1, \dots, b\}$. $b \in [a, c]$ denotes $a \leq b \leq c$ and $b \in (a, c]$ denotes $a < b \leq c$. $Y_{[a:b]}$ denotes the set of random variables $\{Y_i : i = [a : b]\}$ and $Y_{[a,b]}$ denotes the set $\{Y_i : i = a, b\}$. \mathbb{N}^+ denotes the set of positive integers; the function $(x)^+ = \max\{0, x\}$; $\lceil x \rceil$, $\lfloor x \rfloor$ are the ceil, floor functions respectively.

II. SYSTEM MODEL AND PRELIMINARY RESULTS

In this section, we introduce the system model for file storage and delivery in cache-aided systems. We then present achievable schemes for the case of multiple file demands in each transmission interval which are based on repetitions of schemes for single file demands per user.

A. System Model

We consider a cache-aided network (see Fig. 1) with K users and a library of N files, $F_{[1:N]}$, where each file is of size B bits, for $B \in \mathbb{N}^+$. Formally, the files F_n are i.i.d. and distributed as:

$$F_n \sim \text{Unif}\{1, 2, \dots, 2^B\}, \quad \forall n \in [1 : N]. \quad (1)$$

Next, we define the key operational phases and the related performance metric for content storage and delivery in cache-aided systems.

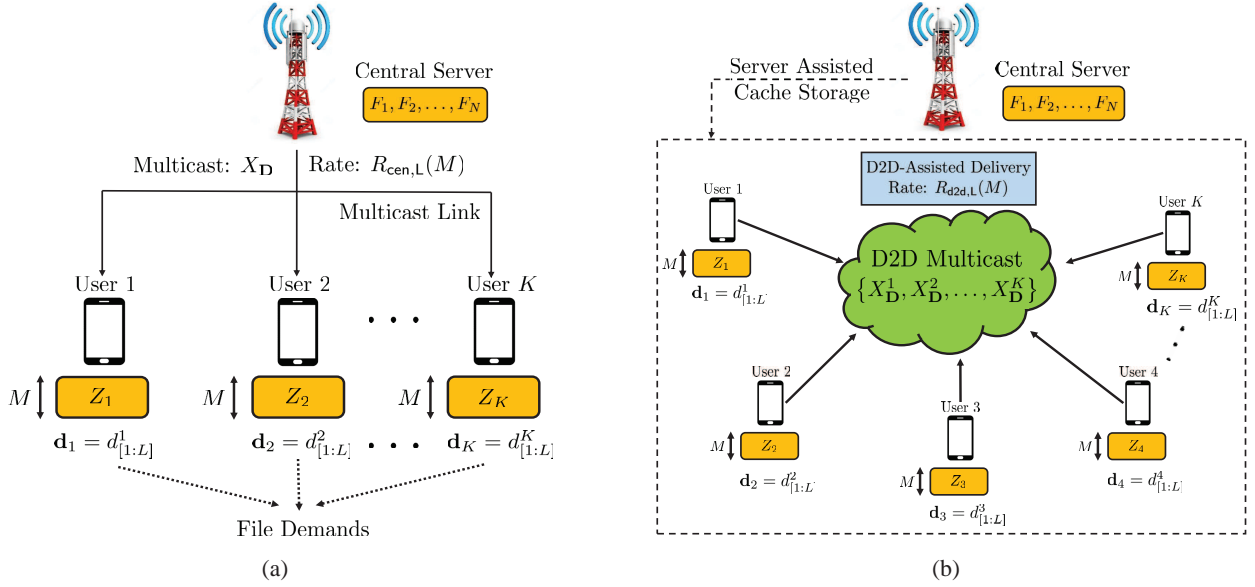


Fig. 1. System Model for cache-aided network with (a) centralized content delivery where the requested content is delivered via multicast transmission by the central server; and (b) D2D-assisted content delivery where each device multicasts to all the other devices using the contents placed in the device cache by the central server.

Definition 1 (Cache Storage). The cache storage phase consists of K caching functions which map the files $F_{[1:N]}$ into the cache content

$$Z_k \triangleq \phi_k(F_{[1:N]}), \quad (2)$$

for each user $k \in [1 : K]$. For cache-aided systems with centralized content delivery the cache storage constraint is such that $H(Z_k) \in [0, MB]^1$. For the case of D2D-assisted delivery, an additional storage constraint is that all caches should be collectively capable of storing the entire library $F_{[1:N]}$ i.e., $KM \geq N$ and $H(Z_k) \in [NB/K, MB]^2$. The cache placement phase generally occurs over a larger time-scale encompassing multiple user demand phases or *transmission intervals*. As a result, the caching functions are agnostic to user demands.

Definition 2 (File Delivery). The file delivery phase occurs in each transmission interval in response to user demands with each user requesting $L \in [1 : N]$ files. The user demands are denoted by $\mathbf{D} = \mathbf{d}_{[1:K]}$, where each users' demand vector consists of L distinct files $\mathbf{d}_k = \mathbf{d}_{[1:L]}^k \in [1 : N]$ for $k \in [1 : K]$. For the case of centralized delivery, the central server uses N^{KL} encoding functions to map the library of files $F_{[1:N]}$ to the multicast transmission

$$X_{\mathbf{D}} \triangleq \psi_{\mathbf{D}}(F_1, \dots, F_N), \quad (3)$$

over the shared link with a rate not exceeding RB bits i.e., $H(X_{\mathbf{D}}) \leq RB$. For D2D-assisted delivery, the encoding function $\psi_{\mathbf{D}}$ is composed of K functions, $\psi_{\mathbf{D}}^k$, one for each user. The K users encode the contents of their respective caches into a composite D2D multicast transmission

$$X_{\mathbf{D}} = \left\{ (X_{\mathbf{D}}^1, X_{\mathbf{D}}^2, \dots, X_{\mathbf{D}}^K) : X_{\mathbf{D}}^k = \psi_{\mathbf{D}}^k(Z_k), \forall k \in [1 : K] \right\}. \quad (4)$$

Each multicast transmission has a rate $\sum_{k=1}^K H(X_{\mathbf{D}}^k) \leq RB$.

Definition 3 (File Decoding). Once the multicast transmission is received, KN^{KL} decoding functions

¹Here $H(Z_k)$ denotes the entropy of the content Z_k stored in the cache of user $k \in [1 : K]$ and represents the total size of Z_k in bits i.e., the cache can store at most M files of size B bits each.

²The lower bound follows from the fact that each cache needs to store at least N/K files.

map the received signal X_D and the local cache content Z_k to the estimates

$$\hat{F}_{d_k} \triangleq \mu_{D,k}(X_D, Z_k), \quad (5)$$

of the L requested files F_{d_k} for user $k \in [1 : K]$. The probability of error is defined as

$$P_e \triangleq \max_{D, k \in [1:K], d \in d_k} \mathbb{P}(\hat{F}_d \neq F_d), \quad (6)$$

i.e., the worst-case probability of error evaluated over all possible demand vectors and across all users for any number of per user demands L .

Definition 4 (*Storage-Rate Tradeoff*). The storage-rate pair $(M, R_{\text{cen},L})$ for centralized delivery or $(M, R_{\text{d2d},L})$ for D2D-assisted delivery is *achievable* if, for any $\epsilon > 0$, there exists a caching and delivery scheme, for which $P_e \leq \epsilon$, where ϵ is an arbitrarily small constant. The optimal storage-rate tradeoffs are defined as

$$R_{\text{cen},L}^*(M) \triangleq \inf \{R_{\text{cen},L} : (M, R_{\text{cen},L}) \text{ is achievable}\}; \quad (7)$$

$$R_{\text{d2d},L}^*(M) \triangleq \inf \{R_{\text{d2d},L} : (M, R_{\text{d2d},L}) \text{ is achievable}\}. \quad (8)$$

B. Preliminary Results

In this section, we present existing achievability results which yield upper bounds on the optimal storage-rate tradeoff for cache-aided systems under centralized as well as D2D-assisted delivery for the case of $L(\geq 1)$ demands per user.

1) *Centralized Delivery with Multiple Demands*: An achievable scheme for caching with centralized delivery was first proposed in [2] for the case of single ($L = 1$) user requests. An extension to the case when each user can make multiple ($L > 1$) demands at any given transmission interval is given by the following lemma.

Lemma 1. For any N files and K users, with each user having cache storage of $M \in \frac{Nt}{K}$ files for any $t \in [0 : K]$, an achievable content delivery rate which upper bounds the optimal rate is given by:

$$R_{\text{cen},L}^*(M) \leq R_{\text{cen},L}(M) = KL \left(1 - \frac{M}{N}\right) \min \left(\frac{1}{1 + KM/N}, \frac{N}{KL} \right), \quad (9)$$

for the case when each user requests any $L \in [1 : N]$ files at every transmission interval.

Proof: The delivery rate in (9) can be achieved by a strategy which repeats L times, the coded multicast delivery scheme proposed in [2, Theorem 1]. The second term inside the $\min(\cdot)$ function is derived from the unicasting of $\min\{N, KL\}$ files. ■

C. D2D-assisted Delivery with Multiple Demands per Device

For the case of D2D-assisted delivery, Ji et. al. proposed an order-optimal caching and delivery scheme in [11] for case of single ($L = 1$) user demands. An extension to the case of multiple ($L > 1$) demands per user, is given by the following lemma.

Lemma 2. For any N files and K users, each having storage size $M \in \frac{Nt}{K}$ files for any $t \in [0 : K]$ with $KM \geq N$, an achievable rate for D2D-assisted content delivery is given by

$$R_{\text{d2d},L}(M) \leq \min \left\{ \frac{LN}{M} \left(1 - \frac{M}{N}\right), N \right\}, \quad (10)$$

for the case when each user requests any $L \in [1 : N]$ files at every transmission interval.

Proof: The delivery rate in (10) can be achieved by a strategy which repeats L times, the coded multicast delivery scheme proposed in [11, Theorem 1]. The second term inside the $\min(\cdot)$ function is derived from the multicasting of all N files, which is possible since the storage constraint for D2D-assisted delivery ensures that $KM \geq N$. ■

In [19], Ji et. al presented a graph-coloring based index coded delivery scheme which showed that coding across files as well as demands can improve the centralized delivery rate compared to the approach in Lemma 1, while D2D-assisted delivery schemes specifically for multiple ($L > 1$) demands has not been studied in literature. In this work, we address the following question - *are the repetition based schemes order-optimal, thereby foregoing the need for more complex approaches?* An answer in the affirmative is provided in Section III, where we leverage the proposed lower bounds to prove the order-optimality of the schemes presented above.

III. MAIN RESULTS AND DISCUSSION

In this section, we present new converse bounds for centralized and D2D-assisted content delivery in cache-aided networks with multiple ($L \geq 1$) demands per user.

A. Centralized Content Delivery

We next present our first main result which gives a new lower bound on the optimal storage-rate tradeoff for cache-aided systems with centralized content delivery.

Theorem 1. *For any N files and K users, each having a cache size of $M \in [0, N]$, the optimal centralized content delivery rate $R_{\text{cen},L}^*(M)$ is lower bounded as*

$$R_{\text{cen},L}^*(M) \geq \max_{\substack{s \in [1: \min\{\lceil N/L \rceil, K\}], \\ \ell \in [1: \lceil N/(Ls) \rceil]}} \frac{1}{\ell} \left\{ N - sM - \frac{\mu(N - L\ell s)^+}{s + \mu} - (N - KL\ell)^+ \right\}, \quad (11)$$

for the case when each user demands $L \in [1 : N]$ files at every transmission interval. The parameter $\mu = (\min(\lceil N/(L\ell) \rceil, K) - s)$, $\forall s, \ell$.

The proof of Theorem 1 is given in Appendix A. The expression in Theorem 1 has two parameters, namely (i) the parameter s , which is related to the number of user caches; and (ii) the parameter ℓ , which is related to multicast transmissions. Compared to the cut-set bounds presented in [19, Theorem 2], the additional parameter ℓ adds further flexibility to the lower bound expression and accounts for file decoding through the interaction of caches and transmissions, yielding a generally tighter lower bound for the case of centralized content delivery with multiple demands per user. The main difference between the cut-set bound and the proposed lower bound is based on the fact that the new bounds better utilize the possible correlation between caches by carefully bounding the joint and conditional entropy of subsets of cache storages by utilizing Han's inequality on subsets (see Section IV-A for more details). The cut-set based lower bound of [19, Theorem 2] is tight only for very large values of cache size M . As shown in the sequel, for such values of M , the proposed bound yields the cut-set bounds for specific choices of s and ℓ and is generally tighter for all other values. This is illustrated in Fig. 2(b) where we show that the proposed bound is strictly tighter than the cut-set bound.

We next present our second main result which shows that an improved approximation of the optimal storage-rate tradeoff can be obtained by use of the proposed lower bound.

Theorem 2. *For any N files and K users, each with a cache size of $M \in [0, N]$, and each user requesting $L(\leq N)$ files at each transmission interval, we have:*

$$\frac{R_{\text{cen},L}(M)}{R_{\text{cen},L}^*(M)} \leq 11. \quad (12)$$

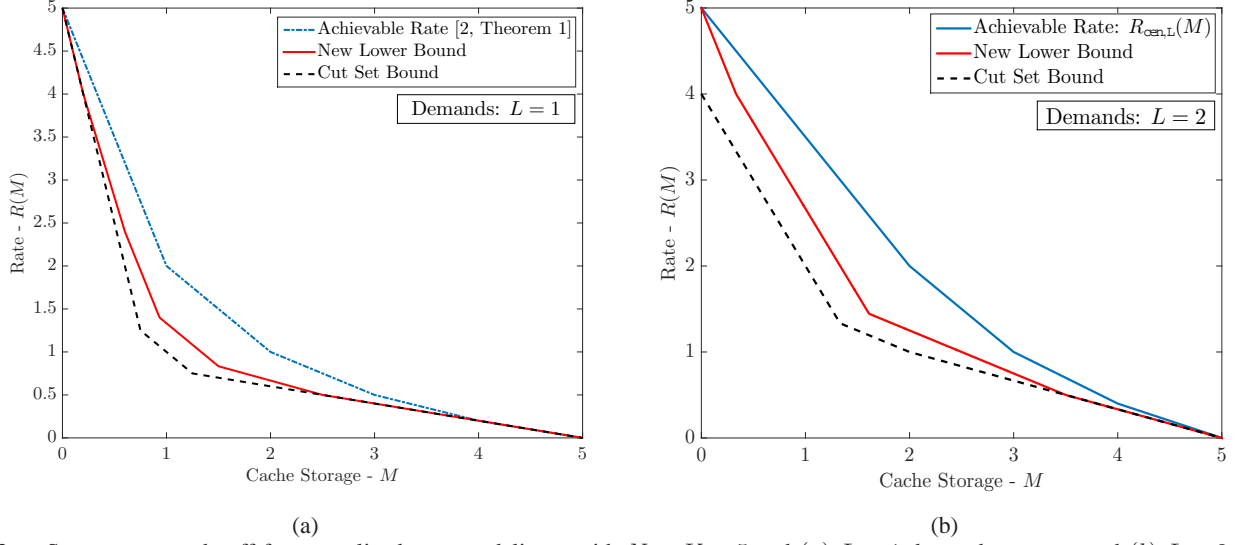


Fig. 2. Storage-rate trade-off for centralized content delivery with $N = K = 5$ and (a) $L = 1$ demand per user; and (b) $L = 2$ demands per user.

The proof of Theorem 2 is provided in Appendix B. This result improves on the gap of 18 between the achievable scheme and the cut-set bound in [19, Theorem 2]. Furthermore, the result shows that performing L repetitions of the scheme for a single demand (as in Lemma 1) is in fact order-optimal, thereby precluding the need for more complex schemes as in [19].

Corollary 1. *For any N files and K users, each having a cache size of $M \in [0, N]$, the optimal centralized content delivery rate $R_{\text{cen}}^*(M)$ for the case when each user requests $L = 1$ file at every transmission interval, is lower bounded by:*

$$R_{\text{cen}}^*(M) \geq \max_{s \in [1:K], \ell \in [1:\lceil N/s \rceil]} \frac{1}{\ell} \left\{ N - sM - \frac{\mu(N - \ell s)^+}{s + \mu} - (N - K\ell)^+ \right\}, \quad (13)$$

where $\mu = (\min(\lceil N/\ell \rceil, K) - s)$, $\forall s, \ell$.

Corollary 1 follows by setting $L = 1$ in Theorem 1 and was originally presented in [22]. The new bounds strictly improve on the cut-set lower bounds presented in [2, Theorem 2] as shown in Fig. 2(a). Using (13), the approximation of the optimal storage-rate tradeoff can be improved as follows.

Theorem 3. *Let $R_{\text{cen}}(M)$ be the achievable rate of the centralized caching scheme given in [2, Theorem 1]. Then, for any K users, N files, and user cache storage in the range $M \in [0, N]$, we have:*

$$\frac{R_{\text{cen}}(M)}{R_{\text{cen}}^*(M)} \leq 8. \quad (14)$$

The proof of Theorem 3 is provided in Appendix C. The result improves on the gap of 12 yielded by the cut-set bound in [2, Theorem 3]³.

B. D2D-Assisted Content Delivery

In this section, we consider the case of D2D-assisted content delivery with each user demanding multiple files in each transmission interval. The next theorem presents our main result which gives the first-known lower bound on the optimal storage-rate tradeoff.

³The results presented in this paper also hold for the case of *decentralized cache placement* as in [3] since the converse makes no assumption on the nature of content placement.

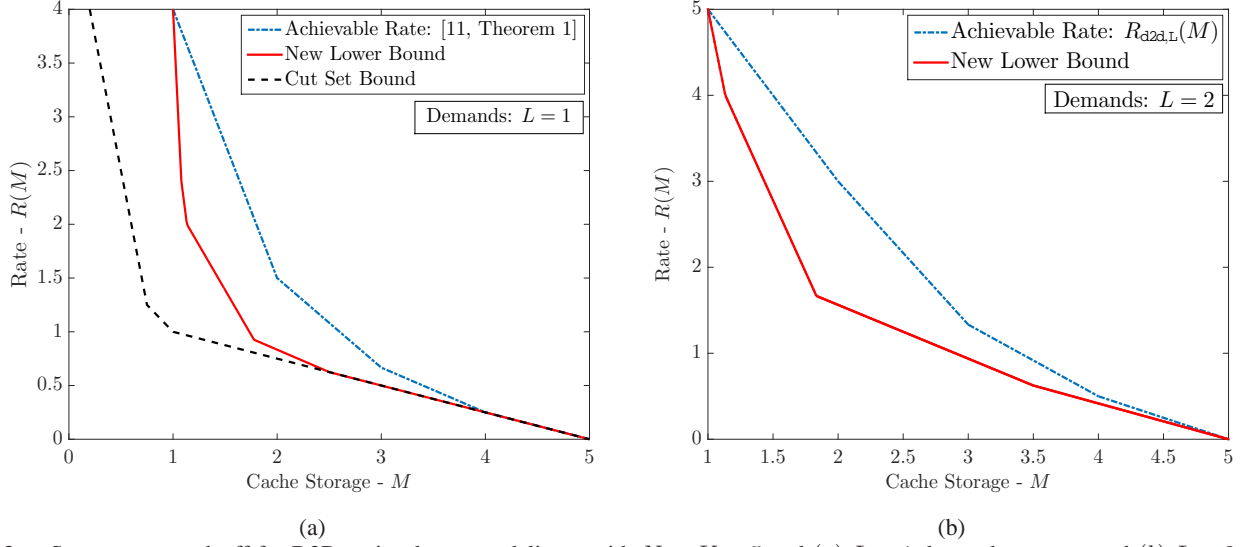


Fig. 3. Storage-rate tradeoff for D2D-assisted content delivery with $N = K = 5$ and (a) $L = 1$ demand per user; and (b) $L = 2$ demands per user.

Theorem 4. For any N files and K users, each having a cache size of $M \in [N/K, N]$, the optimal D2D-assisted content delivery rate $R_{\text{d2d},L}^*(M)$ is lower bounded as

$$R_{\text{d2d},L}^*(M) \geq \max_{s \in [1: \min\{\lceil N/L \rceil, K\}], \ell \in [1: \lceil \frac{N}{Ls} \rceil]} \left\{ \frac{N - sM - \frac{\mu}{s+\mu}(N - L\ell s)^+}{\ell \left(\frac{K-s}{K} \right)} \right\}, \quad (15)$$

for the case when each user demands $L \in [1 : N]$ files at each transmission interval. The parameter $\mu = (\min(\lceil N/(L\ell) \rceil, K) - s)$, $\forall s, \ell$.

The proof of Theorem 4 is presented in Appendix D. Similar to Theorem 1, the parameters s and ℓ yield a family of lower bounds by exploiting the correlation between the caches and transmissions by use of Han's Inequality. Fig. 3(b) shows the lower bound in (15) and the upper bound $R_{\text{d2d},L}(M)$ given in (10). Leveraging the proposed lower bound, we present our second main result in the following theorem.

Theorem 5. For any N files and K users, each having a cache size of $M \in [N/K, N]$, and with each user requesting $L(\leq N)$ files at each transmission interval, we have

$$\frac{R_{\text{d2d},L}(M)}{R_{\text{d2d},L}^*(M)} \leq 10. \quad (16)$$

The proof of Theorem 5 is presented in Appendix E. The result shows that a repetition based scheme as outlined in Lemma 2 is in fact order-optimal and yields a constant factor approximation of the storage-rate trade-off for D2D-assisted content delivery with multiple demands per user.

Corollary 2. For any N files and K users, each having a cache size of $M \in [N/K, N]$, the optimal D2D-assisted content delivery rate $R_{\text{d2d}}^*(M)$, for the case when each user requests $L = 1$ file at every transmission interval, is lower bounded by:

$$R_{\text{d2d}}^*(M) \geq \max_{s \in [1:K], \ell \in [1: \lceil N/s \rceil]} \left\{ \frac{N - sM - \left(\frac{\mu}{s+\mu} \right) (N - \ell s)^+}{\ell \left(\frac{K-s}{K} \right)} \right\}, \quad (17)$$

where $\mu = (\min(\lceil N/\ell \rceil, K) - s)$, $\forall s, \ell$.

Corollary 2 follows by setting $L = 1$ in Theorem 4 and was originally presented in [21]. Compared to the cut-set bound in [11, Theorem 2], we note that the proposed bound in Corollary 2 is always tighter

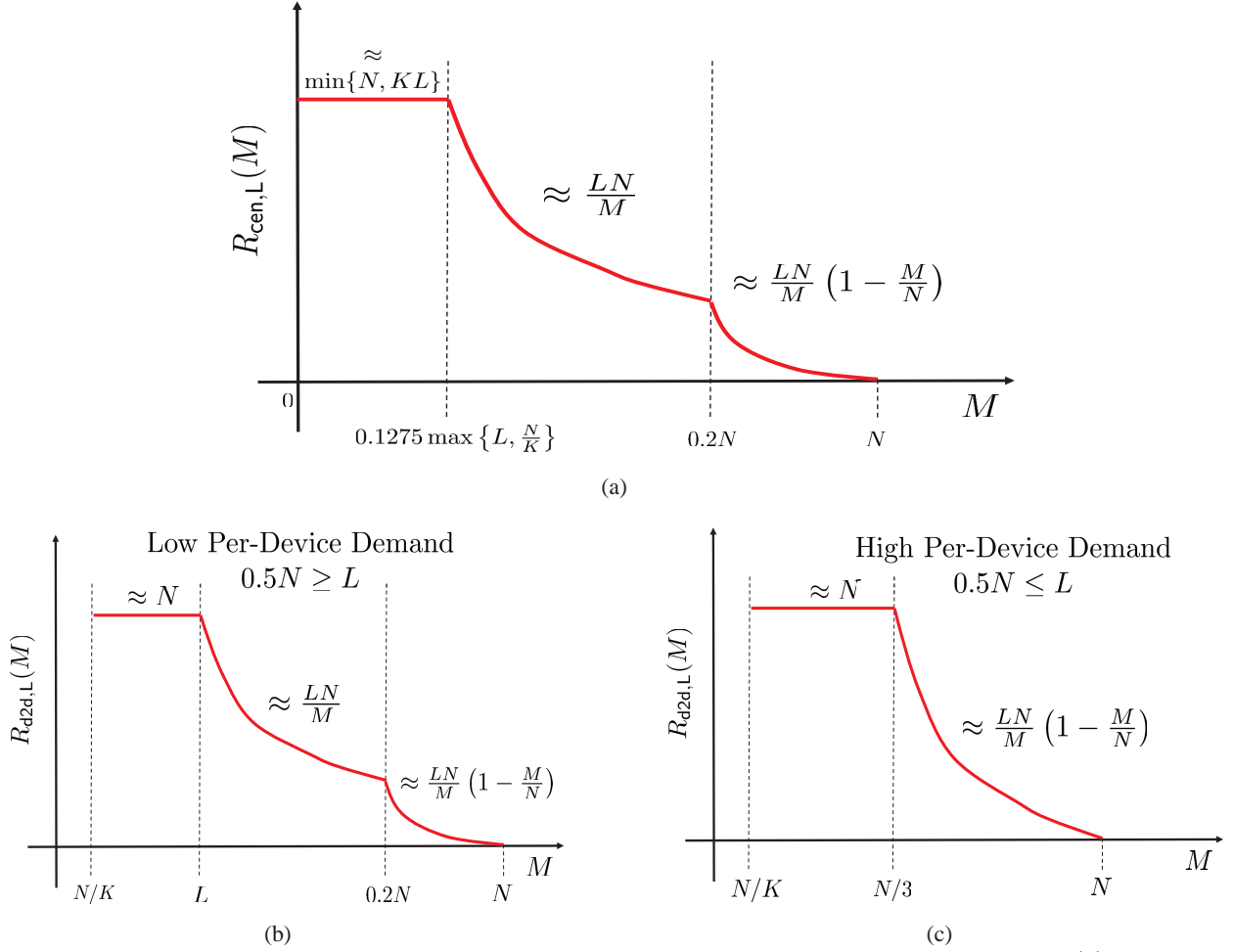


Fig. 4. A representation of the order-optimal approximations to the delivery rate for the repetition based schemes for (a) centralized content delivery, which is used in the proof of Theorem 2; and (b) – (c) for D2D-assisted content delivery with low and high per-device demands, which are used in the proof of Theorem 5.

owing to the additional parameter ℓ and the factor $(K-s)/K \leq 1$ in the denominator of (17). Furthermore, the bound in [11] is tight only for large values of device storage size M . The new bound is tighter for smaller values of M and yields the existing bound as a special case for large values of M .

Using the lower bound in Corollary 2, the optimal storage-rate tradeoff for the case of single demands per user can be approximated as follows.

Theorem 6. For any $K \in \mathbb{N}^+$ user devices, $N \in \mathbb{N}^+$ files, and device storage in the range $M \in [\frac{N}{K}, N]$, we have:

$$\frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \begin{cases} = 1 & M = N/K \\ \leq 3 & M \in (N/K, 2/3] \\ \leq 6 & M \in (2/3, 1] \\ \leq 8 & 1 \leq M \leq N \end{cases}. \quad (18)$$

The proof of Theorem 6 is presented in Appendix F. The result highlights the fact that for the smallest allowable cache size of $M = N/K$, the lower bound in (17) is tight and yields the achievable rate in [11, Theorem 1]. This is also shown in Fig. 3(a) for the case of $N = K = 5$ and $L = 1$.

Remark 1. To prove the order-optimality of the repetition based schemes as shown in Theorems 2 and 5, we use approximations to the achievable rates presented in Lemmas 1 and 2. These approximations are highlighted in Fig. 4. For the case of centralized content delivery, three regimes of cache storage are considered and for very low cache storage, it is approximately optimal to unicast all requested files

as seen in Fig. 4(a). For higher cache storage, a linear dependance of the rate on L/M is established. For the case of D2D-assisted delivery, we see that when users demand less than half the library, three regimes of cache storage need to be considered, while for the case of high per-device demands, only 2 regimes suffice and for storage as high as a third of the library, it is approximately optimal for all users to broadcast all N files from their local caches. Further details are provided in Appendix B and E. \diamond

IV. CASE STUDIES

In this section, we present two case studies to illustrate the new techniques used to obtain the lower bounds in Theorems 1 and 4. For ease of exposition, we consider the special case of $L = 1$ since the results easily extend to any $L > 1$. We show that our technique yields additional bounds as compared to the cut-set techniques in literature and present discussions behind the principal intuitions in applying our method. To this end we first consider the case of centralized content delivery.

A. Centralized Content Delivery: Intuition Behind Proof of Theorem 1

We consider $N = 3$ files, denoted by A, B, C and $K = 3$ users, each with a cache storage M files. For the case of $L = 1$, Corollary 1 yields the following lower bounds for different s, ℓ .

$$\text{New Lower Bounds:} \quad 3R_{\text{cen}}^* + 6M \geq 8, \quad s = 2, \ell = 1; \quad 4R_{\text{cen}}^* + 2M \geq 5, \quad s = 1, \ell = 2 \quad (19)$$

$$\text{Cut-Set Bounds:} \quad R_{\text{cen}}^* + 3M \geq 3, \quad s = 3, \ell = 1; \quad 3R_{\text{cen}}^* + M \geq 3, \quad s = 1, \ell = 3. \quad (20)$$

The existing lower bounds from [2, Theorem 2] are given by (20). The proposed approach provides the additional bounds in (19), thereby yielding tighter lower bounds than [2, Theorem 2] as shown in Fig. 5(a). Next, we detail the derivation of the first bound in (19) highlighting the new aspects and techniques.

To this end, we consider two consecutive requests $(d_1, d_2, d_3) = (A, B, C)$ and $(d_1, d_2, d_3) = (B, C, A)$. It is clear that the first $s = 2$ caches $Z_{[1,2]}$ along with two corresponding transmissions X_{ABC}, X_{BCA} from the central server suffice to decode all the 3 files. We upper bound the entropy of $\ell = 1$ multicast transmission by the optimal rate R_{cen}^* and use the other transmission's decoding capability with the caches to derive the following bound

$$\begin{aligned} 3B &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) \leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) \\ &\leq 2MB + H(X_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) \\ &\stackrel{(a)}{\leq} 2MB + R_{\text{cen}}^* B + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) \\ &\leq 2MB + R_{\text{cen}}^* B + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B) \\ &\leq 2MB + R_{\text{cen}}^* B + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B) + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B) \\ &\stackrel{(b)}{\leq} 2MB + R_{\text{cen}}^* B + H(Z_3|Z_{[1,2]}, A, B) + H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) \\ &\leq 2MB + R_{\text{cen}}^* B + H(Z_3|Z_{[1,2]}, A, B), \end{aligned} \quad (21)$$

where step (a) follows from the fact that $Z_{[1,2]}$ along with X_{ABC} can decode files A, B and step (b) follows from the fact that $H(X_{BCA}|Z_{[1:3]}, X_{ABC}, A, B, C) = 0$ since each transmission is a deterministic function of the files. Considering the term $H(Z_3|Z_{[1,2]}, A, B)$ in (21), we have:

$$H(Z_3|Z_{[1,2]}, A, B) = H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \quad (22)$$

Using (22) in (21), we have:

$$3B \leq 2MB + R_{\text{cen}}^* B + H(Z_{[1:3]}|A, B) - H(Z_{[1,2]}|A, B). \quad (23)$$

Now considering all possible subsets of $Z_{[1:3]}$ with cardinality 2, in the RHS of (23), we have:

$$3B \leq 2MB + R_{\text{cen}}^* B + H(Z_{[1:3]}|A, B) - H(Z_{[2,3]}|A, B) \quad (24)$$

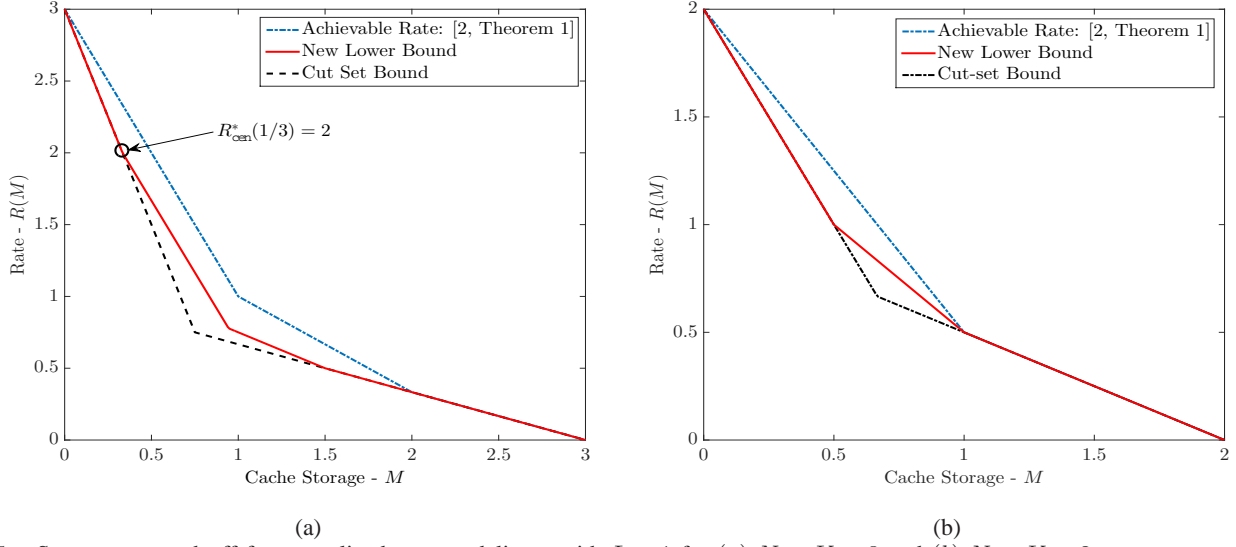


Fig. 5. Storage-rate tradeoff for centralized content delivery with $L = 1$ for (a) $N = K = 3$ and (b) $N = K = 2$.

$$3B \leq 2MB + R_{\text{cen}}^* B + H(Z_{[1:3]}|A, B) - H(Z_{[1,3]}|A, B). \quad (25)$$

Summing (23)-(25), and normalizing by 3, we have:

$$3B \leq 2MB + R_{\text{cen}}^* B + H(Z_{[1:3]}|A, B) - \sum_{i,j=1, i \neq j}^3 \frac{H(Z_{[i,j]}|A, B)}{3}. \quad (26)$$

We next state Han's Inequality [23, Theorem 17.6.1] on subsets of random variables, which we use for further upper bounding (26) in order to derive the proposed lower bound.

Han's Inequality: Let $Y_{[1:n]}$ denote a set of random variables. Further, let $(Y_{[m]}, Y_{[r]}) \subseteq Y_{[1:n]}$ denote subsets of cardinality m, r with $m \leq r$. Han's Inequality states that

$$\frac{1}{\binom{n}{r}} \sum_{Y_{[r]}: |Y_{[r]}|=r} \frac{H(Y_{[r]})}{r} \leq \frac{1}{\binom{n}{m}} \sum_{Y_{[m]}: |Y_{[m]}|=m} \frac{H(Y_{[m]})}{m}, \quad (27)$$

where the sums are over all subsets of cardinality r, m respectively. Next, from (26), consider the set of random variables $Z_{[1:3]}$ and its subsets $(Z_{[1,2]}, Z_{[1,3]}, Z_{[2,3]})$ of cardinality 2. Applying Han's Inequality for these random variables, using $n = r = 3$ and $m = 2$ in (27), we have:

$$\frac{2H(Z_{[1:3]}|A, B)}{3} \leq \sum_{i,j=1, i \neq j}^3 \frac{H(Z_{[i,j]}|A, B)}{3}. \quad (28)$$

Substituting (28) into (26), we have:

$$\begin{aligned} 3B &\leq 2MB + R_{\text{cen}}^* B + H(Z_{[1:3]}|A, B) - \frac{2}{3}H(Z_{[1:3]}|A, B) \\ &\leq 2MB + R_{\text{cen}}^* B + \frac{1}{3}H(Z_{[1:3]}|A, B) \leq 2MB + R_{\text{cen}}^* B + \frac{1}{3}H(Z_{[1:3]}, C|A, B) \\ &\leq 2MB + R_{\text{cen}}^* B + \frac{1}{3} \left(\underbrace{H(C|A, B)}_{\leq 1} + \underbrace{H(Z_{[1:3]}|A, B, C)}_{=0} \right) \leq 2MB + R_{\text{cen}}^* B + \frac{1}{3}B. \end{aligned} \quad (29)$$

Rearranging (29), we get the new lower bound given by the first inequality in (19). The second bound in (19) can be obtained similarly by considering $s = 1$ cache and bounding the entropy of $\ell = 2$ transmissions

by the optimal rate R_{cen}^* and following steps similar to (21)-(29).

Remark 2. We note that the key distinction from the cut-set bounds is the mechanism of bounding the joint entropy of random variables representing the multicast transmissions and the stored contents. Specifically, considering the first inequality in (21), a naive upper bound on the term $H(X_{BCA}|Z_{[1,2]}, X_{ABC})$ would be R_{cen}^* , which would lead to $3 \leq 2M + 2R_{\text{cen}}^*$, which is a loose bound. The main idea is to first observe that given $Z_{[1,2]}$ and the multicast transmission X_{ABC} , the files A, B can be recovered. Hence, we expect a dependence between X_{BCA} and the random variables in the conditioning. In order to capture this dependency, we consider multiple such requests over time, allowing us to write (24), and (25), similar to (23). This symmetrization argument directly leads to the use of Han's inequality and subsequently to the new lower bound. This is the key approach behind Corollary 1 and Theorem 1 which is a general result and holds for all problem parameters. \diamond

Remark 3. Recently [15]–[17] proposed caching and delivery schemes which improve upon the original multicasting scheme presented in [2, Theorem 1]. Specifically, [15] showed that for $K \geq N$, in the small buffer region of $M = 1/K$, the achievable rate is given by $N(1 - M)$ which improves on the achievable rate in [2, Theorem 1]. For $N = K = 3$, the new achievable point $(M, R) = (1/3, 2)$ is highlighted in Fig. 5(a). The lower bound in [2, Th. 2] is shown to be tight only in the regime $0 \leq M \leq 1/K$ for $K \geq N$ in [15]. The lower bound presented in Corollary 1 shows that this is indeed the case and that the new converse is tighter than the cut-set based lower bound for $M > 1/K$ as shown in Fig. 5(a). \diamond

Remark 4. In [2], the authors characterize the optimal storage-rate tradeoff for the case of $N = K = 2$ and show that their lower bound, given by $R_{\text{cen}}^* + 2M \geq 2$ and $2R_{\text{cen}}^* + M \geq 2$, is indeed loose. Our proposed lower bound yields the additional bound, $2R_{\text{cen}}^* + 2M \geq 3$, which makes it tighter than the cut set bound. From Fig. 5(b) and [2] it can be seen that the proposed converse characterizes the optimal rate for the case of $N = K = 2$. \diamond

B. D2D-assisted Content Delivery: Intuition Behind Proof of Theorem 4

We next follow up the discussion in the previous section with an additional example to highlight our proposed techniques for the case of D2D-assisted content delivery with $L = 1$ demand per user. To this end, consider again a system with $N = 3$ files (A, B, C) and $K = 3$ users, each with a cache storage of $M \geq 1$. The proposed lower bound in Corollary 2 gives following bounds for different s, ℓ :

$$\text{New Lower Bounds:} \quad R_{\text{d2d}}^* + 6M \geq 8, \quad s = 2, \ell = 1 \quad (30)$$

$$8R_{\text{d2d}}^* + 6M \geq 15, \quad s = 1, \ell = 2 \quad (31)$$

$$\text{Cut-set Bound:} \quad 2R_{\text{d2d}}^* + M \geq 3, \quad s = 1, \ell = 3, \quad (32)$$

where (32), along with the looser bound $R_{\text{d2d}}^* + 3M \geq 3$, recovers the cut set bound in [19, Theorem 2]. To facilitate the derivation of the new bounds, we first consider the request vectors $(d_1, d_2, d_3) = (A, B, C)$ and $(d_1, d_2, d_3) = (B, C, A)$. The first $s = 2$ cache contents $Z_{[1,2]}$ along with two composite transmissions $X_{ABC} = \{X_{ABC}^3\}$, $X_{BCA} = \{X_{BCA}^3\}$ from the *third* user device are able to decode all 3 files. Here each transmission has the rate of $R_{\text{d2d}}^*/3$. We upper bound the entropy of $\ell = 1$ transmission with this rate and use the other transmission's decoding capability, in conjunction with the cache contents $Z_{[1,2]}$, to derive a tighter bound as follows.

$$\begin{aligned} 3B &\leq H(Z_{[1,2]}, X_{ABC}, X_{BCA}) \leq H(Z_{[1,2]}) + H(X_{ABC}, X_{BCA}|Z_{[1,2]}) \\ &\leq 2MB + H(X_{ABC}) + H(X_{BCA}|Z_{[1,2]}, X_{ABC}) \\ &\leq 2MB + \frac{R_{\text{d2d}}^*}{3}B + H(X_{BCA}|Z_{[1,2]}, X_{ABC}, A, B) \\ &\leq 2MB + \frac{R_{\text{d2d}}^*}{3}B + H(X_{BCA}, Z_3|Z_{[1,2]}, X_{ABC}, A, B) \end{aligned}$$

$$\begin{aligned}
&\leq 2MB + \frac{R_{d2d}^*}{3}B + H(Z_3|Z_{[1,2]}, X_{ABC}, A, B) + H(X_{BCA}|Z_{[1,3]}, X_{ABC}, A, B) \\
&\stackrel{(a)}{\leq} 2MB + \frac{R_{d2d}^*}{3}B + H(Z_3|Z_{[1,2]}, A, B),
\end{aligned} \tag{33}$$

where step (a) follows from the fact that $H(X_{BCA}|Z_{[1,3]}, X_{ABC}, A, B, C) = 0$ since X_{BCA} is a function of the *cache contents* $Z_{[1,3]}$. Considering the term $H(Z_3|Z_{[1,2]}, A, B)$, we have:

$$H(Z_3|Z_{[1,2]}, A, B) = H(Z_{[1,3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{34}$$

Using (34) in (33), we have:

$$3B \leq 2MB + \frac{R_{d2d}^*}{3}B + H(Z_{[1,3]}|A, B) - H(Z_{[1,2]}|A, B). \tag{35}$$

Again, considering all possible subsets of $Z_{[1,3]}$ having cardinality 2, in the RHS of (35), we have

$$3B \leq 2MB + \frac{R_{d2d}^*}{3}B + H(Z_{[1,3]}|A, B) - H(Z_{[2,3]}|A, B). \tag{36}$$

$$3B \leq 2MB + \frac{R_{d2d}^*}{3}B + H(Z_{[1,3]}|A, B) - H(Z_{[1,3]}|A, B). \tag{37}$$

Symmetrizing over the inequalities in (35)-(37), we have:

$$3B \leq 2MB + \frac{R_{d2d}^*}{3}B + H(Z_{[1,3]}|A, B) - \sum_{i,j=1, i \neq j}^3 \frac{H(Z_{[i,j]}|A, B)}{3}. \tag{38}$$

Next, considering the set of caches $Z_{[1,3]}$ and its subsets $Z_{[1,2]}, Z_{[1,3]}Z_{[2,3]}$ of cardinality 2 and applying Han's Inequality (as in (27)), we have from (35)

$$\begin{aligned}
3B &\leq 2MB + \frac{R_{d2d}^*}{3}B + H(Z_{[1,3]}|A, B) - \frac{2H(Z_{[1,3]}|A, B)}{3} \\
&\leq 2MB + \frac{R_{d2d}^*}{3}B + \frac{H(Z_{[1,3]}, C|A, B)}{3} \leq 2MB + \frac{R_{d2d}^*}{3}B + \frac{1}{3}B.
\end{aligned} \tag{39}$$

Rearranging (39), we get the new lower bound in (30). Next, we consider $s = 1$ device cache, Z_1 , and three request vectors $(d_1, d_2, d_3) = (A, B, C)$, $(d_1, d_2, d_3) = (B, C, A)$ and $(d_1, d_2, d_3) = (C, A, B)$ along with the multicast transmissions $X_{ABC} = \{X_{ABC}^2, X_{ABC}^3\}$, $X_{BCA} = \{X_{BCA}^2, X_{BCA}^3\}$, $X_{CAB} = \{X_{CAB}^2, X_{CAB}^3\}$ from users 2, 3, which are capable of decoding all 3 files. In this case, each composite transmission is of rate $2R_{d2d}^*/3$. We upper bound the entropy of $\ell = 2$ transmissions with their rate and following similar steps as the previous case leads us to the lower bound in (31). Finally, considering again, $s = 1$ device storage content, Z_1 , and three request vectors $(d_1, d_2, d_3) = (A, B, C)$, $(d_1, d_2, d_3) = (B, C, A)$ and $(d_1, d_2, d_3) = (C, A, B)$ along with three transmissions $X_{ABC} = \{X_{ABC}^2, X_{ABC}^3\}$, $X_{BCA} = \{X_{BCA}^2, X_{BCA}^3\}$, $X_{CAB} = \{X_{CAB}^2, X_{CAB}^3\}$ which are capable of decoding all 3 files. Each transmission has rate $2R_{d2d}^*/3$. We upper bound the entropy of $\ell = 3$ transmissions by their rates thereby recovering the cut set bound in (32). The new converse is strictly tighter than the cut set bounds. Furthermore, the proposed converse is tight at the point $M = N/K = 1$. Setting $M = 1$ in (30) and comparing with the upper bound from [11, Theorem 1] yields $R_{d2d}^*(1) = 2$ i.e., the achievable scheme proposed in [11] is optimal at $M = 1$.

V. COMPARISONS WITH INDEPENDENT PARALLEL RESULTS

We acknowledge the recent independent contributions from [24]–[28] on developing converse results for cache-aided systems. The authors in [24] derive a new converse bound based on index coding for the case of centralized content delivery with $L = 1$, which shows that the achievable scheme in [2] is optimal if *uncoded* cache placement is assumed. Again, for centralized content delivery, the authors in [25], [26] also obtain improvements over the cut-set bound, for the case when $L = 1$, through different approaches

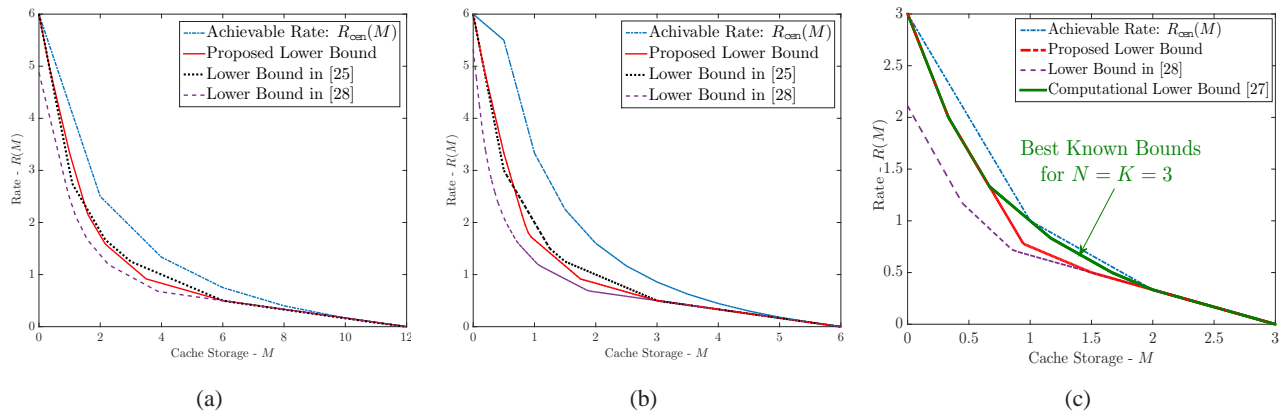


Fig. 6. Comparisons with parallel results for the case of centralized content delivery with $L = 1$ for a cache-aided system with (a) $N = 12, K = 6$; (b) $N = 6, K = 12$ and (c) $N = K = 3$.

than ours. While a direct comparison is analytically intractable, especially owing to the algorithm based approach of [26], we present some numerical comparisons to show that our bounds supersede these bounds in certain regimes of cache storage M for the single demand case. To this end, in Fig. 6(a) and 6(b), we plot the result in [25] which yields the same bound as in [26] for these instances. It can be seen that our bounds are better for the case of low cache memory for both cases. Furthermore, we note that a *holistic lower bound* for centralized content delivery with $L = 1$ is obtained only by combination of all lower bounding approaches in literature and maximizing over the bounds yielded by each method.

The authors in [25] do not derive a constant gap result, however, the authors in [26] show a constant gap of 4 to the achievable rate in [2, Theorem 1]. We emphasize here that the analyses to obtain multiplicative gaps (as in Theorems 2 and 5) are essentially approximations. Thus, deriving lower bounds geared towards tightening this analysis does not guarantee the best known bounds. To this end, we consider the lower bounds presented in [28]. The proposed lower bounds are generally always looser than the cut-set bounds for the case of centralized content delivery with $L = 1$ and by extension than the bounds presented in this paper as shown in Fig. 6. However, the authors leverage the structure of the bounds to approximate the storage-rate tradeoff to within a constant multiplicative factor of 4.7. We note here that the analysis presented in this paper is solely for the purpose of proving the sub-optimality of cut-set bounds in a more general problem setting, i.e., $L \geq 1$, and that the gap to the optimal can be numerically tightened to 3.5 for centralized delivery with $L = 1$, which shows that the bounds are similar to those in [26], [28] in terms of approximately characterizing the optimal storage-rate tradeoff.

Finally, Tian [27] has recently obtained improvements for the specific case of $N = K = 3$ for centralized content delivery with $L = 1$, using a novel computer aided approach as shown in Fig. 6(c). Our proposed method recovers the bound $6M + 3R_{\text{cen}}^* \geq 8$, while the approach in [25], [26] recovers the bound $M + R_{\text{cen}}^* \geq 2$. However, it is unclear whether the bounds $12M + 18R_{\text{cen}}^* \geq 29$ and $3M + 6R_{\text{cen}}^* \geq 8$ can be tractably obtained via analytical methods. Therefore, obtaining the numerical bounds for the $N = K = 3$ system with centralized delivery remains an open problem.

VI. CONCLUSION

In this paper, we presented a new technique for deriving information theoretic lower bounds for cache-aided systems with centralized as well as D2D-assisted content delivery for the general case when users can demand multiple files at each transmission interval. We leveraged Han's Inequality to better model the interaction of user caches and file decoding capabilities of multicast transmissions to derive lower bounds which are strictly tighter than existing cut-set based bounds. Leveraging the proposed lower bounds, we showed that, for the case of multiple demands per user, repeating multiple times, the schemes designed to address content delivery for single demands, is in fact order-optimal for both delivery settings. Furthermore, we provided an approximate characterization of the fundamental storage-rate tradeoff for

centralized content delivery to within a constant multiplicative factor of 11 and for D2D-assisted content delivery to within a factor of 10 for all possible values of problem parameters, thereby improving on the existing results in both paradigms.

APPENDIX A PROOF OF THEOREM 1

Consider a cache-aided system with N files, each of size B bits, and K users, each with a cache size of M files. Let s be an integer such that $s \in [1 : \min\{\lceil N/L \rceil, K\}]$. For the case of centralized delivery with $L \in [1 : N]$ demands per user, the demand vector is such that each user demands L distinct files at each transmission interval. Consider the first s caches $Z_{[1:s]}$ and a demand vector

$$\mathbf{D}_1 = \left(\underbrace{\mathbf{d}_{[1:s]}, \mathbf{d}_{[s+1:K]}}_{= \mathbf{d}_{[1:s]}} = ([1 : L], [L + 1 : 2L], \dots, [L(s - 1) + 1 : Ls], \phi \right), \quad (40)$$

where the first s user demands are for Ls unique files and last $K - s$ users' demands can be for any arbitrary $L(K - s)$ files. To service this set of demands, the central server makes a multicast transmission X_1 , which along with the $Z_{[1:s]}$ is capable of decoding the files $F_{[1:Ls]}$. Similarly, consider another demand,

$$\mathbf{D}_2 = ([Ls + 1 : L(s + 1)], [L(s + 1) + 1 : L(s + 2)], \dots, [L(2s - 1) : 2Ls], \phi), \quad (41)$$

and a resultant multicast transmission X_2 , which along with the s caches, are capable of decoding the files $F_{[Ls+1:2Ls]}$. Thus considering the demand vectors $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{\lceil N/(Ls) \rceil}$ and their corresponding multicast transmissions $X_1, X_2, \dots, X_{\lceil N/(Ls) \rceil}$, along with the first s caches $Z_{[1:s]}$, the whole library of files $F_{[1:N]}$ can be decoded. Considering $B = 1$ without loss of generality. We have:

$$\begin{aligned} N &\leq I(F_{1:N}; Z_{[1:s]}, X_{[1:\lceil N/(Ls) \rceil]}) \leq H(Z_{[1:s]}, X_{[1:\lceil N/(Ls) \rceil]}) \leq H(Z_{[1:s]}) + H(X_{[1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}) \\ &\leq sM + H(X_{[1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}) \leq sM + H(X_{[1:\ell]} | Z_{[1:s]}) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}, X_{[1:\ell]}) \\ &\stackrel{(a)}{\leq} sM + \ell R_{\text{cen},L}^*(M) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}) \\ &\stackrel{(b)}{\leq} sM + \ell R_{\text{cen},L}^*(M) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]}, Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}) \\ &\leq sM + \ell R_{\text{cen},L}^*(M) + \underbrace{H(Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]})}_{\triangleq \delta} \\ &\quad + \underbrace{H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell s]})}_{\triangleq \lambda}, \end{aligned} \quad (42)$$

where step (a) results from bounding the entropy of $\ell \in \{1, 2, \dots, \lceil N/(Ls) \rceil\}$ transmissions given the caches $Z_{[1:s]}$ by $\ell R_{\text{cen},L}^*(M)$, where each transmission is of rate $R_{\text{cen},L}^*(M)$. Furthermore, the caches $Z_{[1:s]}$ with transmissions $X_{[1:\ell]}$ can decode files $F_{[1:L\ell s]}$. In step (b), μ number of caches are introduced into the entropy, where μ is the number of remaining caches which along with caches $Z_{[1:s]}$ and transmissions $X_{[1:\ell]}$, can decode the remaining $(N - L\ell s)$ files. It is to be noted that all the remaining $K - s$ caches might not be required for decoding all files. Thus we have:

$$\mu = \min \left\{ \left\lceil \frac{N - L\ell s}{L\ell} \right\rceil, K - s \right\} = \min \{ \lceil N/(L\ell) \rceil, K \} - s, \quad (43)$$

where the last equality follows since s is an integer. Next, we obtain upper bounds on the two terms δ and λ in (42).

Upper Bound on δ : We consider the factor δ , from (42) and upper bound it as follows:

$$\delta = H(Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}) \leq H(Z_{[s+1:s+\mu]} | Z_{[1:s]}, F_{[1:L\ell s]})$$

$$= H(Z_{[1:s+\mu]}|F_{[1:L\ell s]}) - H(Z_{[1:s]}|F_{[1:L\ell s]}). \quad (44)$$

Considering all possible subsets of $Z_{[1:s+\mu]}$ having cardinality s , i.e., considering all possible combinations of distinct files in the request vectors and all possible combinations of s caches in (42), we can obtain $\binom{s+\mu}{s}$ different inequalities of the form of (44). Symmetrizing over all the inequalities, we have:

$$\delta \leq H(Z_{[1:s+\mu]}|F_{[1:L\ell s]}) - \sum_{i=1}^{\binom{s+\mu}{s}} \frac{H(Z_{[s]}^i|F_{[1:L\ell s]})}{\binom{s+\mu}{s}}, \quad (45)$$

where, $Z_{[s]}^i$ is the i -th subset of $Z_{[1:s+\mu]}$ with cardinality s . Next, consider $Z_{[1:s+\mu]}$ as the set of random variables $\{Z_k : k \in 1, \dots, s + \mu\}$ and the subsets $Z_{[s]}^i \subseteq Z_{[1:s+\mu]}$, $\forall i = 1, \dots, \binom{s+\mu}{s}$. Applying Han's Inequality from (27), we have:

$$\frac{s}{s+\mu} H(Z_{[1:s+\mu]}|F_{[1:L\ell s]}) \leq \frac{1}{\binom{s+\mu}{s}} \sum_{i=1}^{\binom{s+\mu}{s}} H(Z_{[s]}^i|F_{[1:L\ell s]}). \quad (46)$$

Substituting (46) into (45), we have:

$$\begin{aligned} \delta &\leq H(Z_{[1:s+\mu]}|F_{[1:L\ell s]}) - \frac{s}{s+\mu} H(Z_{[1:s+\mu]}|F_{[1:L\ell s]}) \\ &= \frac{\mu}{s+\mu} H(Z_{[1:s+\mu]}|F_{[1:L\ell s]}) \leq \frac{\mu}{s+\mu} H(Z_{[1:s+\mu]}, F_{[L\ell s+1:N]}|F_{[1:L\ell s]}) \\ &= \frac{\mu}{s+\mu} \left(H(F_{[L\ell s+1:N]}|F_{[1:L\ell s]}) + \underbrace{H(Z_{[1:s+\mu]}|F_{[1:N]})}_{=0} \right) \stackrel{(a)}{\leq} \frac{\mu}{s+\mu} (N - L\ell s)^+, \end{aligned} \quad (47)$$

where step (a) follows from the fact that the caches are functions of all N files in the library.

Upper Bound on λ : To upper bound λ , we observe from the last step in (42) that the transmissions $X_{[1:\ell]}$, along with caches $Z_{[1:s+\mu]}$ can decode the files $F_{[1:L\ell(s+\mu)]}$ within the conditioning, i.e.,

$$\lambda = H(X_{[\ell+1:\lceil N/(Ls) \rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell(s+\mu)]}). \quad (48)$$

In order to characterize the upper bound on λ , we consider two cases as follows.

• **Case 1 ($N \leq L\ell(s + \mu)$) :** All files are decoded by the caches $Z_{[1:s+\mu]}$ and transmissions $X_{[1:\ell]}$ within the conditioning for the term λ in (42). We have

$$\lambda = H(X_{[\ell+1:\lceil N/(Ls) \rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}) = 0, \quad (49)$$

since all transmissions are functions of the file library $F_{[1:N]}$. In the case when, for $N > K$, fewer than K caches suffices to decode all files with the transmissions within the conditioning in λ i.e. $s + \mu \leq K$, we have:

$$KL\ell \geq L\ell(s + \mu) \geq N, \quad \text{i.e.,} \quad \lambda = (N - KL\ell)^+ = 0. \quad (50)$$

It can also be easily seen that for the case of $K \geq N$, $\lambda = (N - KL\ell)^+ = 0$ since $\ell, L \geq 1$.

• **Case 2 ($N > L\ell(s + \mu)$) :** The case when, even with $s + \mu = K$ caches, all files are not decoded by the caches and transmissions within the conditioning for the term λ in (42). In this case, $\lambda \neq 0$ and we have:

$$\begin{aligned} \lambda &= H(X_{[\ell+1:\lceil N/(Ls) \rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}) \\ &\leq H(X_{[\ell+1:\lceil N/(Ls) \rceil]}, F_{[KL\ell+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}) \\ &\leq H(F_{[KL\ell+1:N]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:KL\ell]}) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]}|Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:N]}) \end{aligned}$$

$$\stackrel{(a)}{\leq} H(F_{[KL\ell+1:N]}) \leq (N - KL\ell), \quad (51)$$

where step (a) follows from the fact that the second entropy term in the previous step goes to zero since transmissions are functions of the N files. Thus from (49) and (51), we can compactly bound λ as:

$$\lambda \leq (N - KL\ell)^+. \quad (52)$$

Substituting (47) and (52) into (42), we have:

$$N \leq sM + \ell R_{\text{cen,L}}^*(M) + \frac{\mu}{s + \mu}(N - L\ell s)^+ + (N - KL\ell)^+ \quad (53)$$

Rearranging (53), we obtain the following lower bound on the optimal rate $R_{\text{cen,L}}^*(M)$

$$R_{\text{cen,L}}^*(M) \geq \frac{1}{\ell} \left\{ N - sM - \frac{\mu}{s + \mu}(N - L\ell s)^+ - (N - KL\ell)^+ \right\}. \quad (54)$$

Optimizing over all parameter values of s, ℓ , completes the proof of Theorem 1.

APPENDIX B PROOF OF THEOREM 2

From Theorem 1, considering the lower bound on the optimal rate $R_{\text{cen,L}}^*(M)$, we set $\ell = \lceil \frac{\beta N}{Ls} \rceil \in [1 : \lceil \frac{N}{Ls} \rceil]$ with $\beta \in [0, 1]$. Using this, we next derive an upper bound on the term $\left(\frac{\mu}{\mu+s}\right)$ as follows

$$\begin{aligned} \frac{\mu}{\mu + s} &= \frac{\min \left\{ \lceil \frac{N}{L\ell} \rceil, K \right\} - s}{\min \left\{ \lceil \frac{N}{L\ell} \rceil, K \right\}} \leq 1 - \frac{s}{\lceil \frac{N}{L\ell} \rceil} = 1 - \frac{s}{\left\lceil \frac{N}{L \lceil \frac{\beta N}{Ls} \rceil} \right\rceil} \leq 1 - \frac{s}{\left\lceil \frac{s}{\beta} \right\rceil} \\ &\leq 1 - \frac{s}{\frac{s}{\beta} + 1} = 1 - \frac{\beta}{1 + \frac{\beta}{s}} \leq 1 - \frac{\beta}{1 + \beta} = \frac{1}{1 + \beta}, \end{aligned} \quad (55)$$

where the last inequality follows from the fact that $s \geq 1$. Substituting (55) into (11), we have:

$$\begin{aligned} R_{\text{cen,L}}^*(M) &\geq \frac{N - sM - \frac{1}{1+\beta} (N - L \lceil \frac{\beta N}{Ls} \rceil s)^+ - (N - KL \lceil \frac{\beta N}{Ls} \rceil)^+}{\lceil \frac{\beta N}{Ls} \rceil} \\ &\geq \frac{\left(\frac{2\beta}{1+\beta}\right) N - sM - N \left(1 - K \frac{\beta}{s}\right)^+}{\lceil \frac{\beta N}{Ls} \rceil}. \end{aligned} \quad (56)$$

Next, we consider two cases, namely (i) $\min \left\{ \frac{N}{L}, K \right\} \leq 10$; and (ii) $\min \left\{ \frac{N}{L}, K \right\} \geq 11$.

• **Case 1** ($\min \left\{ \frac{N}{L}, K \right\} \leq 10$): For this case, setting $s = 1$ and $\beta = 1$ in (56), we have the following form on the lower bound,

$$R_{\text{cen,L}}^*(M) \geq \frac{N \left(1 - \frac{M}{N}\right)}{\left\lceil \frac{N}{L} \right\rceil} \quad (57)$$

Consider first, the case when $\frac{N}{L} \leq K$. From (9), we have the following upper bound on the achievable rate

$$R_{\text{cen,L}}(M) \leq \min\{N, KL\} \left(1 - \frac{M}{N}\right) \leq N \left(1 - \frac{M}{N}\right). \quad (58)$$

Therefore, we have

$$\text{Gap} = \frac{R_{\text{cen,L}}(M)}{R_{\text{cen,L}}^*(M)} \leq \left\lceil \frac{N}{L} \right\rceil \leq 10. \quad (59)$$

Next, consider the case when $K \leq \frac{N}{L}$. Again, from (9), we have the following upper bound on the achievable rate

$$R_{\text{cen,L}}(M) \leq \min\{N, KL\} \left(1 - \frac{M}{N}\right) \leq KL \left(1 - \frac{M}{N}\right). \quad (60)$$

Again, setting $s = 1$ and $\beta = 1$ in (56), we have

$$R_{\text{cen,L}}^*(M) \geq \frac{N \left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} = \frac{L \left(1 - \frac{M}{N}\right)}{1 + \frac{L}{N}} \geq \frac{KL \left(1 - \frac{M}{N}\right)}{1 + K} \quad (61)$$

Therefore, we have

$$\text{Gap} = \frac{R_{\text{cen,L}}(M)}{R_{\text{cen,L}}^*(M)} \leq K + 1 \leq 10 + 1 = 11. \quad (62)$$

• **Case 2** ($\min\{\frac{N}{L}, K\} \geq 11$) : For this case, we consider three distinct regimes for the cache storage size M : *Regime 1*: $0 \leq M \leq 1.275 \max\{L, N/K\}$; *Regime 2*: $1.275 \max\{L, N/K\} < M \leq 0.2N$; and *Regime 3*: $0.2N < M \leq N$. We consider each of the three regimes separately.

• **Regime 1** ($0 \leq M \leq 1.275 \max\{L, N/K\}$) :

For this regime, we set $s = \lfloor 0.3049 \min\{N/L, K\} \rfloor \in [1 : \min\{N/L, K\}]$ and $\ell = \lceil \frac{0.9649N}{Ls} \rceil$, from (56), we have

$$\begin{aligned} R_{\text{cen,L}}^*(M) &\geq \frac{\left(\frac{2 \times 0.9649}{1+0.9649}\right) - s \frac{M}{N} - \left(1 - K \frac{0.9649}{s}\right)^+}{\frac{0.9649}{Ls} + \frac{1}{N}} \\ &= \frac{\left(\frac{2 \times 0.9649}{1+0.9649}\right) - \lfloor 0.3049 \min\{N/L, K\} \rfloor \frac{1.275 \max\{L, N/K\}}{N} - \left(1 - K \frac{0.9649}{\lfloor 0.3049 \min\{N/L, K\} \rfloor}\right)^+}{\frac{0.9649}{L \lfloor 0.3049 \min\{N/L, K\} \rfloor} + \frac{1}{N}} \\ &\stackrel{(a)}{\geq} \frac{\left(\frac{2 \times 0.9649}{1+0.9649}\right) - (0.3049 \times 1.275) \frac{\min\{N/L, K\} \max\{L, N/K\}}{N} - \left(1 - K \frac{0.9649}{0.3049 \min\{N/L, K\}}\right)^+}{\frac{0.9649}{L(0.3049 \min\{N/L, K\} - 1)} + \frac{1}{N}} \\ &\geq \frac{L \min\left\{\frac{N}{L}, K\right\} \left(0.3049 - \frac{1}{\min\left\{\frac{N}{L}, K\right\}}\right) \left\{\left(\frac{2 \times 0.9649}{1+0.9649}\right) - (0.3049 \times 1.275) - \left(1 - \frac{0.9649}{0.3049}\right)^+\right\}}{0.9649 + \frac{L(0.3049 \min\left\{\frac{N}{L}, K\right\} - 1)}{N}} \\ &\stackrel{(b)}{\geq} \frac{\min\{N, KL\} \left(0.3049 - \frac{1}{10+1}\right) \left\{\left(\frac{2 \times 0.9649}{1+0.9649}\right) - (0.3049 \times 1.275) - \left(1 - \frac{0.9649}{0.3049}\right)^+\right\}}{0.9649 + 0.3049} \\ &\geq \frac{\min\{N, KL\}}{10}, \end{aligned} \quad (63)$$

where step (a) follows by using $\lfloor 0.3049 \min\{N/L, K\} \rfloor \leq 0.3049 \min\{N/L, K\}$ in the numerator and $\lfloor 0.3049 \min\{N/L, K\} \rfloor \geq 0.3049 \min\{N/L, K\} - 1$ in the denominator; and step (b) follows by using $\min\{N/L, K\} \leq N/L$ in the second term in the denominator. Again, considering the upper bound in (9), we have

$$R_{\text{cen,L}}(M) \leq \min\{N, KL\} \left(1 - \frac{M}{N}\right) \leq \min\{N, KL\}. \quad (64)$$

Therefore for *Regime 1*, we have

$$\text{Gap} = \frac{R_{\text{cen,L}}(M)}{R_{\text{cen,L}}^*(M)} \leq 10. \quad (65)$$

• **Regime 2** ($1.275 \max\{L, N/K\} < M \leq 0.2N$) :

For this regime, setting $s = \lfloor 0.442 \frac{N}{M} \rfloor \in [1 : \min\{N/L, K\}]^4$ and $\ell = \lceil \frac{0.984N}{Ls} \rceil$, from (56), we have

$$\begin{aligned}
R_{\text{cen,L}}^*(M) &\geq \frac{\left(\frac{2 \times 0.984}{1+0.984}\right) - s \frac{M}{N} - \left(1 - K \frac{0.984}{s}\right)^+}{\frac{0.984}{Ls} + \frac{1}{N}} \\
&= \frac{\left(\frac{2 \times 0.984}{1+0.984}\right) - \lfloor 0.442 \frac{N}{M} \rfloor \frac{M}{N} - \left(1 - K \frac{0.984}{\lfloor 0.442 \frac{N}{M} \rfloor}\right)^+}{\frac{0.984}{L \lfloor 0.442 \frac{N}{M} \rfloor} + \frac{1}{N}} \\
&\stackrel{(a)}{\geq} \frac{\left(\frac{2 \times 0.984}{1+0.984}\right) - 0.442 \frac{N}{M} \frac{M}{N} - \left(1 - \frac{0.984}{0.442} \frac{KM}{N}\right)^+}{\frac{0.984}{L(0.442 \frac{N}{M} - 1)} + \frac{1}{N}} \\
&\stackrel{(b)}{\geq} \frac{\frac{LN}{M} (0.442 - \frac{M}{N}) \left\{ \left(\frac{2 \times 0.984}{1+0.984}\right) - 0.442 - \left(1 - \frac{0.984}{0.442} \times 1.275\right)^+ \right\}}{0.984 + \frac{0.442}{1.275} \frac{L}{M}} \\
&\stackrel{(c)}{\geq} \frac{\frac{LN}{M} (0.442 - 0.2) \left\{ \left(\frac{2 \times 0.984}{1+0.984}\right) - 0.442 - \left(1 - \frac{0.984 \times 1.275}{0.442}\right)^+ \right\}}{0.984 + \frac{0.442}{1.275}} \geq \frac{LN}{10M}, \tag{66}
\end{aligned}$$

where step (a) follows again by using $\lfloor 0.442N/M \rfloor \leq 0.442N/M$ in the numerator and $\lfloor 0.442N/M \rfloor \geq 0.442N/M - 1$ in the denominator; step (b) follows from using $KM/N \geq 1.275$; and step (c) follows by using $M/N \leq 0.2$ in the numerator and $M \geq 1.275L$ in the denominator. Again considering the upper bound in (9), we have

$$R_{\text{cen,L}}(M) \leq \frac{KL \left(1 - \frac{M}{N}\right)}{1 + \frac{KM}{N}} \leq \frac{LN}{M} \left(1 - \frac{M}{N}\right) \leq \frac{LN}{M}. \tag{67}$$

Therefore for *Regime 2*, we have

$$\text{Gap} = \frac{R_{\text{cen,L}}(M)}{R_{\text{cen,L}}^*(M)} \leq 10. \tag{68}$$

• **Regime 3** ($0.2N < M \leq N$) :

In this regime, setting $s = 1$ and $\ell = \lceil \frac{N}{L} \rceil$ in (56), we note that in this case, $\mu = 0$ and $(N - K\ell)^+ = 0$. Thus, we have

$$R_{\text{cen,L}}^*(M) \geq \frac{N \left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} \geq \frac{\left(1 - \frac{M}{N}\right)}{\frac{1}{L} + \frac{1}{N}}. \tag{69}$$

From (9), we have

$$R_{\text{cen,L}}(M) \leq \frac{KL \left(1 - \frac{M}{N}\right)}{1 + \frac{KM}{N}} \leq \frac{LN}{M} \left(1 - \frac{M}{N}\right). \tag{70}$$

Therefore for *Regime 3*, we have

$$\text{Gap} = \frac{R_{\text{cen,L}}(M)}{R_{\text{cen,L}}^*(M)} \leq \frac{LN}{M} \left(\frac{1}{L} + \frac{1}{N}\right) \leq \frac{2N}{M} \leq \frac{2}{0.2} \leq 10 \tag{71}$$

Combining (59),(62),(65),(68) and (71), completes the proof of Theorem 2. \square

⁴The range of s is validated as follows. Using the upper bound $M \leq 0.2N$, we have $0.442N/M \geq 0.442/0.2 \geq 1$. Again using the lower bound $M \geq 1.275L$, we have $0.442N/M \leq \frac{0.442}{1.275} N/L \leq N/L$. Again using $M \geq 1.275N/K$, we have $0.442N/M \leq K$.

APPENDIX C
PROOF OF THEOREM 3

From Corollary 1, considering the lower bound on the optimal rate $R_{\text{cen}}^*(M)$, we set $\ell = \lceil \frac{\beta N}{s} \rceil \in \{1, 2, \dots, \lceil \frac{N}{s} \rceil\}$ with $0 < \beta \leq 1$. Using this we next derive an upper bound on $\left(\frac{\mu}{\mu+s}\right)$.

$$\begin{aligned} \frac{\mu}{\mu+s} &= \frac{\min\left\{\lceil \frac{N-\ell s}{\ell} \rceil, K-s\right\}}{\min\left\{\lceil \frac{N-\ell s}{\ell} \rceil, K-s\right\}+s} = \frac{\min\left\{\lceil \frac{N}{\ell} \rceil, K\right\}-s}{\min\left\{\lceil \frac{N}{\ell} \rceil, K\right\}} \\ &= 1 - \frac{s}{\min\left\{\lceil \frac{N}{\ell} \rceil, K\right\}} = 1 - \frac{s}{\min\left\{\left\lceil \frac{N}{\lceil \frac{\beta N}{s} \rceil} \right\rceil, K\right\}} \\ &\leq 1 - \frac{s}{\left\lceil \frac{s}{\beta} \right\rceil} \leq 1 - \frac{s}{\frac{s}{\beta} + 1} = 1 - \frac{\beta}{1 + \frac{\beta}{s}} \leq 1 - \frac{\beta}{1 + \beta} = \frac{1}{1 + \beta}, \end{aligned} \quad (72)$$

where the last inequality follows from the fact that $s \geq 1$. Substituting (72) into (13), we have:

$$\begin{aligned} R_{\text{cen}}^*(M) &\geq \frac{N - sM - \frac{1}{1+\beta} (N - \lceil \frac{\beta N}{s} \rceil s)^+ - (N - K \lceil \frac{\beta N}{s} \rceil)^+}{\lceil \frac{\beta N}{s} \rceil} \\ &\geq \frac{N - sM - N \left(\frac{1-\beta}{1+\beta}\right) - N (1 - K \frac{\beta}{s})^+}{\lceil \frac{\beta N}{s} \rceil}. \end{aligned} \quad (73)$$

Next, we consider two cases, namely (i) $\min\{N, K\} \leq 8$; and (ii) $\min\{N, K\} \geq 9$. We next consider each case separately.

• **Case 1 ($\min\{N, K\} \leq 8$)** : For this case, setting $s = 1$ and $\beta = 1$ in (73), we have:

$$R_{\text{cen}}^*(M) \geq \frac{N - M}{N} = \left(1 - \frac{M}{N}\right). \quad (74)$$

Again, from [2, Theorem 1], we have:

$$R_{\text{cen}}(M) \leq \min\{N, K\} \left(1 - \frac{M}{N}\right). \quad (75)$$

Thus for this case, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\text{cen}}(M)}{R_{\text{cen}}^*(M)} \leq \min\{N, K\} \leq 8. \quad (76)$$

• **Case 2 ($\min\{N, K\} \geq 9$)** : For this case, we consider three distinct regimes for the cache storage size M namely (i) *Regime 1*: $0 \leq M \leq 1.01 \max\{1, N/K\}$; (ii) *Regime 2*: $1.01 \max\{1, N/K\} < M \leq 0.1250N$; and (iii) *Regime 3*: $0.1250N < M \leq N$. We next consider each of the three regimes separately.

• **Regime 1 ($0 \leq M \leq 1.01 \max\{1, N/K\}$)** :

In this regime, setting $s = \lfloor 0.4701 \min\{N, K\} \rfloor \in \{1, 2, \dots, \min\{N, K\}\}$, $\ell = \lceil \frac{0.93N}{s} \rceil$, and using the fact that $x \leq \lceil x \rceil \leq x + 1$ and $x - 1 \leq \lfloor x \rfloor \leq x$, we have:

$$R_{\text{cen}}^*(M) \geq \frac{N - sM - N \left(\frac{1-\beta}{1+\beta}\right) - N (1 - \frac{K\beta}{s})^+}{\frac{\beta N}{s} + 1} \geq \frac{N \left[\frac{2\beta}{1+\beta} - s \frac{M}{N} - (1 - \frac{K\beta}{s})^+\right]}{\frac{\beta N}{s} + 1}$$

$$\begin{aligned}
& \geq \frac{\left\{ \frac{2 \times 0.93}{1+0.93} - \lfloor 0.4701 \min\{N, K\} \rfloor \frac{M}{N} - \left(1 - \frac{0.93K}{\lfloor 0.4701 \min\{N, K\} \rfloor} \right)^+ \right\}}{\frac{0.93}{\lfloor 0.4701 \min\{N, K\} \rfloor} + \frac{1}{N}} \\
& \geq \frac{\left\{ \frac{2 \times 0.93}{1+0.93} - 0.4701 \min\{N, K\} \frac{1.01 \max\{1, N/K\}}{N} - \left(1 - \frac{0.93K}{0.4701 \min\{N, K\}} \right)^+ \right\}}{\frac{0.93}{0.4701 \min\{N, K\} - 1} + \frac{1}{N}} \\
& \geq \frac{\left\{ (0.4701 \min\{N, K\} - 1) \left[\frac{2 \times 0.93}{1+0.93} - 0.4701 \times 1.01 - \left(1 - \frac{0.93}{0.4701} \right)^+ \right] \right\}}{0.93 + \frac{0.4701 \min\{N, K\}}{N} - \frac{1}{N}} \\
& \geq \min\{N, K\} \frac{(0.4701 - \frac{1}{9})}{0.93 + 0.4701} \left[\frac{2 \times 0.93}{1 + 0.93} - 0.4701 \times 1.01 - \left(1 - \frac{0.93}{0.4701} \right)^+ \right] \\
& \geq \frac{\min\{N, K\}}{8}.
\end{aligned} \tag{77}$$

Again, from [2, Theorem 1], we have:

$$R_{\text{cen}}(M) \leq \min\{N, K\} \left(1 - \frac{M}{N}\right) \leq \min\{N, K\}. \tag{78}$$

Thus for this regime, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\text{cen}}(M)}{R_{\text{cen}}^*(M)} \leq 8. \tag{79}$$

• **Regime 2** ($1.01 \max\{1, N/K\} < M \leq 0.1250N$):

In this regime, we set $s = \lfloor 0.4983 \frac{N}{M} \rfloor \in \{1, 2, \dots, \min\{N, K\}\}$, $\ell = \lceil \frac{0.991N}{s} \rceil$ and using the fact that $x \leq \lceil x \rceil \leq x + 1$ and $x - 1 \leq \lfloor x \rfloor \leq x$, we have:

$$\begin{aligned}
R_{\text{cen}}^*(M) & \geq \frac{N \left[\frac{2\beta}{1+\beta} - s \frac{M}{N} - \left(1 - \frac{K\beta}{s}\right)^+ \right]}{\frac{\beta N}{s} + 1} \geq \frac{N \left[\frac{2 \times 0.991}{1+0.991} - 0.4983 - \left(1 - \frac{0.991}{0.4983} \frac{KM}{N}\right)^+ \right]}{\frac{0.991N}{0.4983 \frac{N}{M} - 1} + 1} \\
& \geq \frac{N \left[\frac{2 \times 0.991}{1+0.991} - 0.4983 - \left(1 - \frac{0.991 \times 1.01}{0.4983}\right)^+ \right]}{\frac{0.991N}{0.4983 \frac{N}{M} - 1} + 1} \\
& \geq \frac{(0.4983 \frac{N}{M} - 1) \left[\frac{2 \times 0.991}{1+0.991} - 0.4983 - \left(1 - \frac{0.991 \times 1.01}{0.4983}\right)^+ \right]}{0.991 + 0.4983 \frac{1}{M} - \frac{1}{N}} \\
& \geq \frac{N}{M} \frac{(0.4983 - \frac{M}{N}) \left[\frac{2 \times 0.991}{1+0.991} - 0.4983 - \left(1 - \frac{0.991 \times 1.01}{0.4983}\right)^+ \right]}{0.991 + \frac{0.4983}{1.01}} \\
& \geq \frac{N}{M} \frac{(0.4983 - 0.1250)}{0.991 + \frac{0.4983}{1.01}} \left[\frac{2 \times 0.991}{1 + 0.991} - 0.4983 - \left(1 - \frac{0.991 \times 1.01}{0.4983}\right)^+ \right] \geq \frac{N}{8M}.
\end{aligned} \tag{80}$$

Again, from [2, Theorem 1], we have:

$$R_{\text{cen}}(M) \leq \frac{\min\{N, K\}}{1 + \frac{KM}{N}} \left(1 - \frac{M}{N}\right) \leq \frac{K}{\frac{KM}{N}} \left(1 - \frac{M}{N}\right)$$

$$\leq \frac{N}{M} \left(1 - \frac{M}{N}\right) \leq \frac{N}{M}. \quad (81)$$

Thus for this regime, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\text{cen}}(M)}{R_{\text{cen}}^*(M)} \leq 8. \quad (82)$$

• **Regime 3** ($0.1250N < M \leq N$) :

In this regime, setting $s = 1$ and $\beta = 1$ i.e., $\ell = N$, we have:

$$R_{\text{cen}}^*(M) \geq \frac{N - M}{N} = \left(1 - \frac{M}{N}\right). \quad (83)$$

Again, from [2, Theorem 1], we have:

$$\begin{aligned} R_{\text{cen}}(M) &\leq \frac{\min\{N, K\}}{1 + \frac{KM}{N}} \left(1 - \frac{M}{N}\right) \leq \frac{K}{\frac{KM}{N}} \left(1 - \frac{M}{N}\right) \\ &\leq \frac{N}{M} \left(1 - \frac{M}{N}\right) \leq \frac{1}{0.1250} \left(1 - \frac{M}{N}\right). \end{aligned} \quad (84)$$

Thus for this regime, the gap between the upper and lower bound is given by:

$$\text{Gap} = \frac{R_{\text{cen}}(M)}{R_{\text{cen}}^*(M)} \leq \frac{1}{0.1250} = 8. \quad (85)$$

Thus from (76), (79), (82) and (85), we have for all N, K , the gap between the achievability and the proposed converse is upper bounded by 8. This completes proof of Theorem 3. \square

APPENDIX D PROOF OF THEOREM 4

Consider the case of D2D-assisted content delivery for cache-aided system with a library of $N \in \mathbb{N}^+$ files $F_{[1:N]}$ each of size B bits, and $K \in \mathbb{N}^+$ users, with cache storage $Z_{[1:K]}$ which satisfies the minimum D2D storage constraint $KM \geq N$. Let s be an integer such that $s \in [1 : \min\{\lceil N/L \rceil, K\}]$. The demand vector is such that each user requests L distinct files at each transmission interval. Consider the first s caches $Z_{[1:s]}$ and a demand vector

$$\mathbf{D}_1 = (\mathbf{d}_{[1:s]}, \mathbf{d}_{[s+1:K]}) = \left(\underbrace{\{1 : L\}, \{L+1 : 2L\}, \dots, \{L(s-1)+1 : Ls\}}_{= \mathbf{d}_{[1:s]}}, \phi \right), \quad (86)$$

where the first s user demands are for Ls unique files and last $K - s$ users' demands can be for any arbitrary $L(K - s)$ files. To service this set of demands, consider a composite multicast transmission

$$X_1 = \left\{ X_{(\mathbf{d}_{[1:s]}, \phi)}^{s+1}, X_{(\mathbf{d}_{[1:s]}, \phi)}^{s+2}, \dots, X_{(\mathbf{d}_{[1:s]}, \phi)}^K \right\}, \quad (87)$$

composed of $(K - s)$ multicast transmissions, which, along with the s device caches decodes the files $F_{[1:Ls]}$. Similarly consider another demand vector,

$$\mathbf{D}_2 = \left(\{Ls+1 : L(s+1)\}, \{L(s+1)+1 : L(s+2)\}, \dots, \{L(2s-1) : 2Ls\}, \phi \right). \quad (88)$$

A second composite multicast transmission X_2 , along with device cache contents $Z_{[1:s]}$, can decode the next Ls files $F_{[Ls+1:2Ls]}$. Thus considering the request vectors $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{\lceil N/(Ls) \rceil}$ and their corresponding composite multicast transmissions $X_1, X_2, \dots, X_{\lceil N/(Ls) \rceil}$, along with the first s device caches $Z_{[1:s]}$, the whole library of files $F_{[1:N]}$ can be decoded. Note that for an optimal composite transmission rate $R_{\text{d2d}}^*(M)$,

each device in the D2D cluster multicasts with a rate of $R_{\text{d2d}}^*(M)/K$. Thus the entropy of each composite transmission, consisting of $K - s$ transmissions, can be upper bounded as

$$H(X_i) \leq \frac{(K-s)}{K} R_{\text{d2d,L}}^*(M), \quad \forall i \in [1 : \lceil N/(Ls) \rceil] \quad (89)$$

Considering $B = 1$ without loss of generality, we have the following sequence of inequalities

$$\begin{aligned} N &\leq I(F_{1:N}; Z_{[1:s]}, X_{[1:\lceil N/(Ls) \rceil]}) \leq H(Z_{[1:s]}, X_{[1:\lceil N/(Ls) \rceil]}) \leq H(Z_{[1:s]}) + H(X_{[1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}) \\ &\leq sM + H(X_{[1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}) \leq sM + H(X_{[1:\ell]} | Z_{[1:s]}) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}, X_{[1:\ell]}) \\ &\stackrel{(a)}{\leq} sM + \frac{\ell(K-s)}{K} R_{\text{d2d,L}}^*(M) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}) \\ &\stackrel{(b)}{\leq} sM + \frac{\ell(K-s)}{K} R_{\text{d2d,L}}^*(M) + H(X_{[\ell+1:\lceil N/(Ls) \rceil]}, Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]}) \\ &\leq sM + \frac{\ell(K-s)}{K} R_{\text{d2d,L}}^*(M) + \underbrace{H(Z_{[s+1:s+\mu]} | Z_{[1:s]}, X_{[1:\ell]}, F_{[1:L\ell s]})}_{\triangleq \delta} \\ &\quad + \underbrace{H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s+\mu]}, X_{[1:\ell]}, F_{[1:L\ell s]})}_{\triangleq \lambda}, \end{aligned} \quad (90)$$

where step (a) follows from (89) and that the device storage contents, $Z_{[1:s]}$, along with the composite transmission vectors $X_{[1:\ell]}$ are capable of decoding the files $F_{[1:L\ell s]}$. In step (b), $\mu = (\min\{\lceil N/(L\ell) \rceil, K\} - s)$ is the number of additional device caches which, along with the transmissions $X_{[1:\ell]}$ can decode all N files. Note that, for $s = K$, we have:

$$H(X_{[1:\lceil N/(Ls) \rceil]} | Z_{[1:s]}) = 0, \quad (91)$$

since transmissions are functions of all K caches. As a result, the second step in (90) yields the minimum storage constraint for D2D-assisted delivery $KM \geq N$. Next we upper bound the terms δ, λ in (90) which finally yields an upper bound on the RHS. We first note that the term δ is identical to the case of centralized delivery and can be upper bounded using Han's Inequality by following the same steps as in (44)-(47) in Appendix A, yielding the upper bound

$$\delta \leq \frac{\mu}{s+\mu} (N - L\ell s)^+. \quad (92)$$

Upper Bound on λ : We next derive an upper bound on the factor λ in (90) and consider two distinct cases as follows.

• **Case 1 ($N \leq L\ell(s + \mu)$) :** We consider the case that all N files can be decoded with $\mu \leq K - s$ additional device storage contents and transmissions $X_{[1:\ell]}$, within the conditioning in the factor λ in (90), i.e., $L\ell(s + \mu) \geq N$. Thus, we have

$$\lambda = H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:s+\mu]}, F_{[1:N]}) = 0, \quad (93)$$

which follows from the fact that the transmissions are functions of all N files.

• **Case 2 ($N > L\ell(s + \mu)$) :** We consider the complementary case where $\mu = K - s$ additional device storage contents along with the transmissions $X_{[1:\ell]}$, cannot decode all N files. We have:

$$\lambda = H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:K]}, F_{[1:KL\ell]}) \leq H(X_{[\ell+1:\lceil N/(Ls) \rceil]} | Z_{[1:K]}) = 0, \quad (94)$$

which follows from the fact that $KM \geq N$ i.e., all files are stored within the collective device caches for D2D-assisted delivery and hence all transmissions are functions of the cache contents. Thus combining (93) and (94) we have:

$$\lambda = 0. \quad (95)$$

Substituting (92) and (95) into (90) and optimizing over all parameter values of s, ℓ , completes the proof of Theorem 4. \square

APPENDIX E PROOF OF THEOREM 5

From Theorem 4, considering the lower bound on the optimal rate $R_{\text{d2d,L}}^*(M)$, we set $\ell = \lceil \frac{\beta N}{Ls} \rceil \in [1 : \lceil \frac{N}{Ls} \rceil]$ with $\beta \in [0, 1]$. We make use of the upper bound on $\left(\frac{\mu}{\mu+s}\right)$ from (55) in Appendix B. Using this in (15) from Theorem 4, we have

$$R_{\text{d2d,L}}^*(M) \geq \frac{N - sM - \frac{1}{1+\beta} (N - L \lceil \frac{\beta N}{Ls} \rceil s)^+}{\lceil \frac{\beta N}{Ls} \rceil \left(\frac{K-s}{K}\right)} \geq \frac{N \left(\frac{2\beta}{1+\beta} - s \frac{M}{N}\right)}{\lceil \frac{\beta N}{Ls} \rceil \left(\frac{K-s}{K}\right)} \quad (96)$$

In order to facilitate the proof of Theorem 5, we consider two cases namely - (i) *low per-device demand* with $0.5N \geq L$; and (ii) *high per-device demand* with $0.5N \leq L$. We consider the two cases separately.

• **Case 1 ($0.5N \geq L$):** For the case of low-per device demands, we divide the available cache storage at each device into the following three regimes, namely (i) *Regime 1*: $N/K \leq M \leq L$; (ii) *Regime 2*: $L \leq M \leq 0.2N$; and (iii) *Regime 3*: $0.2N \leq M \leq N$. We consider each regime separately.

• **Regime 1 ($N/K \leq M \leq L$):**

For this regime of cache storage, we consider two further sub-cases, i.e., (i) $N < K$ and (ii) $N \geq K$. We next treat each of the sub-cases separately.

– **Sub-case 1 ($N < K$):** For this sub-case, we note that from the minimum storage constraint for D2D-assisted delivery, i.e., $KM \geq N$, the minimum allowable cache storage at each user can be less than unity. Therefore, we divide the available cache storage in this regime into two sub-regimes namely (i) $N/K \leq M \leq 0.5$ and (ii) $0.5 \leq M \leq L$. We treat these sub-regimes separately as follows. Consider first, the sub-regime i.e., $N/K \leq M \leq 0.5$. For this sub-regime consider the case when $N = 1$. For this case, setting $s = 1$ and $\beta = 1$, from the lower bound in (96), we have

$$R_{\text{d2d,L}}^*(M) \geq (1 - M), \quad (97)$$

where we have used the fact that $L = 1$ when $N = 1$. Again considering the upper bound in (10), we have $R_{\text{d2d,L}} \leq 1$. Using the upper and the lower bounds, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}}{R_{\text{d2d,L}}^*} \leq \frac{1}{1 - M} \leq \frac{1}{1 - 0.5} = 2. \quad (98)$$

Next, we consider the case when $N \geq 2$. For this case, setting $s = \lceil \frac{N}{L} \rceil \in [1 : \lceil \frac{N}{L} \rceil]$ and $\beta = 1$, from (96), we have

$$\begin{aligned} R_{\text{d2d,L}}^*(M) &\geq \frac{N \left(1 - \lceil \frac{N}{L} \rceil \frac{M}{N}\right)}{\left\lceil \frac{N}{L \lceil \frac{N}{L} \rceil} \right\rceil \frac{K - \lceil \frac{N}{L} \rceil}{K}} \geq N \left(1 - \left(\frac{N}{L} + 1\right) \frac{M}{N}\right) \\ &= N \left(1 - \left(\frac{1}{L} + \frac{1}{N}\right) M\right) \stackrel{(a)}{\geq} N \left(1 - \frac{3}{2} \times 0.5\right), \end{aligned} \quad (99)$$

where step (a) follows from the fact that $N \geq 2$ and $L \geq 1$. Again, from the upper bound in (10), we have $R_{\text{d2d,L}} \leq N$. Using this, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}}{R_{\text{d2d,L}}^*} \leq \frac{1}{1 - \frac{3}{2} \times 0.5} = 4. \quad (100)$$

We next consider the sub-regime $0.5 \leq M \leq L$. In this regime, setting $s = \lfloor 0.5 \frac{N}{M} \rfloor \in [1 : K]^5$ and $\beta = 1$, from the lower bound in (96), we have

$$R_{\text{d2d,L}}^* \geq \frac{N(1 - \lfloor 0.5 \frac{N}{M} \rfloor \frac{M}{N})}{\left\lceil \frac{N}{L \lfloor 0.5 \frac{N}{M} \rfloor} \right\rceil \frac{K - \lfloor 0.5 \frac{N}{M} \rfloor}{K}} \geq \frac{N(1 - 0.5)}{\left\lceil \frac{N/L}{\lfloor 0.5 \frac{N}{L} \rfloor} \right\rceil} \stackrel{(a)}{\geq} \frac{N(1 - 0.5)}{3}, \quad (101)$$

where step (a) follows from the fact that for any $N/L \geq 2$, we have $\frac{N/L}{\lfloor 0.5(N/L) \rfloor} \leq 3$. Again from the upper bound in (10), we have $R_{\text{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}(M)}{R_{\text{d2d,L}}^*(M)} \leq \frac{3}{1 - 0.5} = 6. \quad (102)$$

– **Sub-case 2 ($N \geq K$):** For this sub-case, we note that from the minimum storage constraint for D2D-assisted delivery, i.e., $KM \geq N$, we have $M \geq 1$. Therefore, we consider the following regime of available cache storage $0.5 \leq N/K \leq M \leq L$. In this regime, setting $s = \lfloor 0.5 \frac{N}{M} \rfloor \in [1 : K]^6$ and $\beta = 1$, from the lower bound in (96), we have

$$R_{\text{d2d,L}}^* \geq \frac{N(1 - \lfloor 0.5 \frac{N}{M} \rfloor \frac{M}{N})}{\left\lceil \frac{N}{L \lfloor 0.5 \frac{N}{M} \rfloor} \right\rceil \frac{K - \lfloor 0.5 \frac{N}{M} \rfloor}{K}} \geq \frac{N(1 - 0.5)}{\left\lceil \frac{N/L}{\lfloor 0.5 \frac{N}{L} \rfloor} \right\rceil} \stackrel{(a)}{\geq} \frac{N(1 - 0.5)}{3}, \quad (103)$$

where step (a) again follows from the fact that for any $N/L \geq 2$, we have $\frac{N/L}{\lfloor 0.5(N/L) \rfloor} \leq 3$. Again from the upper bound in (10), we have $R_{\text{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}(M)}{R_{\text{d2d,L}}^*(M)} \leq \frac{3}{1 - 0.5} = 6. \quad (104)$$

• **Regime 2 ($L \leq M \leq 0.2N$):**

For this regime, setting $s = \lfloor 0.51 \frac{N}{M} \rfloor \in [1 : K]^7$ and $\ell = \lceil \frac{0.984N}{Ls} \rceil$, from (96), we have

$$\begin{aligned} R_{\text{d2d,L}}^*(M) &\geq \frac{\left(\frac{2 \times 0.984}{1+0.984}\right) - s \frac{M}{N}}{\left(\frac{0.984}{Ls} + \frac{1}{N}\right) \left\lceil \frac{K-s}{K} \right\rceil} \stackrel{(a)}{\geq} \frac{\left(\frac{2 \times 0.984}{1+0.984}\right) - \lfloor 0.51 \frac{N}{M} \rfloor \frac{M}{N}}{\frac{0.984}{L \lfloor 0.51 \frac{N}{M} \rfloor} + \frac{1}{N}} \stackrel{(b)}{\geq} \frac{\left(\frac{2 \times 0.984}{1+0.984}\right) - 0.51 \frac{N}{M} \frac{M}{N}}{\frac{0.984}{L(0.51 \frac{N}{M} - 1)} + \frac{1}{N}} \\ &\geq \frac{\frac{LN}{M} (0.51 - \frac{M}{N}) \left\{ \left(\frac{2 \times 0.984}{1+0.984}\right) - 0.51 \right\}}{0.984 + 0.51 \left(\frac{L}{M} - \frac{L}{N}\right)} \stackrel{(c)}{\geq} \frac{\frac{LN}{M} (0.51 - 0.2) \left\{ \left(\frac{2 \times 0.984}{1+0.984}\right) - 0.51 \right\}}{0.984 + 0.51} \geq \frac{LN}{10M}, \end{aligned} \quad (105)$$

where step (a) follows due to the fact that $(K-s)/K \leq 1$; step (b) follows by using $\lfloor 0.51N/M \rfloor \leq 0.51N/M$ in the numerator and $\lfloor 0.51N/M \rfloor \leq 0.51N/M - 1$ in the denominator; and step (c) follows by using $M/N \leq 0.2$ in the numerator and $M \geq L$ in the denominator. Again, considering the upper bound in (10), we have

$$R_{\text{d2d,L}}(M) \leq \frac{LN}{M} \left(1 - \frac{M}{N}\right) \leq \frac{LN}{M}. \quad (106)$$

⁵The regime of s can be verified as follows. Using the lower bound $0.5 \leq M$, we have $0.5N/M \leq N < K$. Again using the upper bound $M \leq L$, we have $0.5N/M \geq 0.5N/L \geq 1$.

⁶The regime of s is validated as follows. Using the lower bound $M \geq N/K$, we have $0.5N/M \leq 0.5K \leq K$. Again using the upper bound $M \leq L$, we have $0.5N/M \geq 0.5N/L \geq 1$.

⁷The regime of s can be validated as follows. Consider first, a lower bound on $0.5N/M$. In the given regime, we have $0.5N/M \geq 0.5/0.2 \geq 1$. Next, we consider an upper bound on $0.5N/M$. Consider first, the case when $N/L \leq K$. In this case, it's easy to note that $0.5N/M \leq K$. Next consider the case that $N/L \geq K$. In this case, *Regime 2* reduces to $L \leq N/K \leq M \leq 0.2N$ due to the minimum storage constraint and hence we have $0.5N/M \leq 0.5K \leq K$. Therefore we have $\lfloor 0.5N/M \rfloor \in [1 : K]$.

Therefore for *Regime 2*, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}(M)}{R_{\text{d2d,L}}^*(M)} \leq 10. \quad (107)$$

• **Regime 3** ($0.2N \leq M \leq N$):

In this regime, setting $s = 1$ and $\beta = 1$ in (96), we have

$$R_{\text{d2d,L}}^*(M) \geq \frac{N \left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} \geq \frac{\left(1 - \frac{M}{N}\right)}{\frac{1}{L} + \frac{1}{N}}. \quad (108)$$

Again, considering the upper bound in (10), we have

$$R_{\text{d2d,L}}(M) \leq \frac{LN}{M} \left(1 - \frac{M}{N}\right). \quad (109)$$

Therefore for *Regime 3*, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}(M)}{R_{\text{d2d,L}}^*(M)} \leq \frac{LN}{M} \left(\frac{1}{L} + \frac{1}{N}\right) \stackrel{(a)}{\leq} \frac{LN}{M} \times \frac{2}{L} \leq \frac{2N}{M} \leq \frac{2}{0.2} \leq 10, \quad (110)$$

where step (a) follows from the fact that $L \leq N$.

• **Case 2** ($0.5N \leq L$): For the case of high per-device demands, we divide the available cache storage at each device into the following two regimes, namely (i) *Regime 1*: $N/K \leq M \leq N/3$; and (ii) *Regime 2*: $N/3 \leq M \leq N$. We next consider each regime separately.

• **Regime 1** ($N/K \leq M \leq N/3$):

For this regime, setting $s = 1$ and $\ell = \lceil \frac{0.5N}{Ls} \rceil$, from (96), we have

$$R_{\text{d2d,L}}^*(M) \geq \frac{N \left(\left(\frac{2 \times 0.5}{1+0.5}\right) - \frac{M}{N}\right)}{\lceil \frac{0.5N}{L} \rceil \left(\frac{K-1}{K}\right)} \geq \frac{N \left(\left(\frac{2 \times 0.5}{1+0.5}\right) - \frac{M}{N}\right)}{\frac{0.5N}{L} + 1} \stackrel{(a)}{\geq} \frac{N \left(\left(\frac{2 \times 0.5}{1+0.5}\right) - \frac{1}{3}\right)}{2}, \quad (111)$$

where step (a) follows by using the lower bound $L \geq 0.5N$. Again, considering the upper bound in (10), we have $R_{\text{d2d,L}}(M) \leq N$. Using the upper and lower bounds, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}(M)}{R_{\text{d2d,L}}^*(M)} \leq \frac{2}{\left(\frac{2 \times 0.5}{1+0.5}\right) - \frac{1}{3}} \leq 6. \quad (112)$$

• **Regime 2** ($N/3 \leq M \leq N$):

In this regime, setting $s = 1$ and $\beta = 1$ in (96), we have

$$R_{\text{d2d,L}}^*(M) \geq \frac{N \left(1 - \frac{M}{N}\right)}{\frac{N}{L} + 1} \geq \frac{\left(1 - \frac{M}{N}\right)}{\frac{1}{L} + \frac{1}{N}}. \quad (113)$$

From (10), we have

$$R_{\text{d2d,L}}(M) \leq \frac{LN}{M} \left(1 - \frac{M}{N}\right). \quad (114)$$

Therefore for *Regime 3*, we have

$$\text{Gap} = \frac{R_{\text{d2d,L}}(M)}{R_{\text{d2d,L}}^*(M)} \leq \frac{LN}{M} \left(\frac{1}{L} + \frac{1}{N}\right) \leq \frac{LN}{M} \times \frac{2}{L} \leq \frac{2N}{M} \leq \frac{2}{1/3} = 6 \quad (115)$$

Finally, combining (98), (100), (102), (104), (107), (110), (112) and (115), completes the proof of Theorem 5. \square

APPENDIX F
PROOF OF THEOREM 6

From Corollary 2, considering the lower bound on the optimal rate $R_{\text{d2d}}^*(M)$, we set $\ell = \lceil \frac{\beta N}{s} \rceil \in \{1, 2, \dots, \lceil \frac{N}{s} \rceil\}$ with $0 < \beta \leq 1$. Under this setting, we can again use the upper bound on $\left(\frac{\mu}{\mu+s}\right)$ from (72) from Appendix C. Substituting (72) into (17), we have the following form on the lower bound

$$R_{\text{d2d}}^*(M) \geq \frac{N - sM - \frac{1}{1+\beta} (N - \lceil \frac{\beta N}{s} \rceil s)^+}{\lceil \frac{\beta N}{s} \rceil \left(\frac{K-s}{K}\right)} \geq \frac{N \left(\frac{2\beta}{1+\beta} - s\frac{M}{N}\right)}{\lceil \frac{\beta N}{s} \rceil \left(\frac{K-s}{K}\right)}. \quad (116)$$

In order to facilitate the proof of Theorem 6, we consider the three cases, namely (i) the gap at $M = N/K$; (ii) the case when $N < K$; and (iii) the case when $N \geq K$. We next consider each of these cases separately.

• **Case 1 (Gap at $M = N/K$):** In this case, setting $s = N$ and $\beta = 1$ in (116), we have

$$R_{\text{d2d}}^*(M) \geq \frac{N(1-M)}{\left(\frac{K-N}{K}\right)} = \frac{KN(1-\frac{N}{K})}{K-N} = N. \quad (117)$$

Again, from [11, Theorem 1], we have

$$R_{\text{d2d}}(M) \leq \frac{N}{M} \left(1 - \frac{M}{N}\right) \leq N, \quad (118)$$

where the last inequality stems from the fact that for $N \geq K$, $M \geq 1$. Thus the gap is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq 1. \quad (119)$$

• **Case 2 ($N < K$):** For $N < K$ we divide the cache memory into 3 distinct regimes: *Regime 1*: $N/K < M \leq 1$; *Regime 2*: $1 < M \leq 0.1250N$; and *Regime 3*: $0.1250N < M \leq N$. We consider each of the regimes separately.

• **Regime 1 ($N/K < M \leq 1$):**

For this regime, we consider two sub-regimes, namely (i) $N/K < M \leq 2/3$; and (ii) $2/3 < M \leq 1$. We consider each sub-regime separately.

– **Sub-regime 1 ($N/K < M \leq 2/3$):** From (116), setting $s = N$ and $\beta = 1$, we have

$$R_{\text{d2d}}^*(M) \geq \frac{N(1-M)}{\left(\frac{K-N}{K}\right)} \geq N(1-M) \geq N \left(1 - \frac{2}{3}\right) = \frac{N}{3} \quad (120)$$

Again from [11, Theorem 1], we have

$$R_{\text{d2d}} \triangleq \min \left\{ \frac{N}{M} \left(1 - \frac{M}{N}\right), N \right\} \leq N \quad (121)$$

Thus in this sub-regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq \frac{N}{N/3} \leq 3. \quad (122)$$

– **Sub-regime 2 ($2/3 < M \leq 1$):** Consider first, the case when $N = 1$. Setting $s = 1$, $\beta = 1$ in (116), we have

$$R_{\text{d2d}}^*(M) \geq \frac{1-sM}{\lceil 1/s \rceil} \geq 1-M. \quad (123)$$

From [11, Theorem 1], setting $N = 1$, we have

$$R_{\text{d2d}}(M) \leq \frac{1}{M}(1 - M). \quad (124)$$

Thus the gap in this case is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq \frac{1}{M} \leq \frac{3}{2}. \quad (125)$$

For the case of $N \geq 2$, setting $s = \lfloor \frac{2N}{3M} \rfloor$ and $\beta = 1$ in (116), we have

$$R_{\text{d2d}}^*(M) \geq \frac{N \left(1 - \lfloor \frac{2N}{3M} \rfloor \frac{M}{N}\right)}{\left\lfloor \frac{N}{\lfloor \frac{2N}{3M} \rfloor} \right\rfloor \left(\frac{K-s}{K}\right)} \geq \frac{N \left(1 - \lfloor \frac{2N}{3M} \rfloor \frac{M}{N}\right)}{\left\lfloor \frac{N}{\lfloor \frac{2N}{3M} \rfloor} \right\rfloor} \geq \frac{N \left(1 - \frac{2}{3}\right)}{\lfloor 2 \rfloor} \geq \frac{N}{6}, \quad (126)$$

where, we have used the fact that $\frac{N}{\lfloor 2N/3 \rfloor} \leq 2$, $\forall N \geq 2$. Again from [11, Theorem 1], we have

$$R_{\text{d2d}} \triangleq \min \left\{ \frac{N}{M} \left(1 - \frac{M}{N}\right), N \right\} \leq N \quad (127)$$

Thus in this sub-regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq \frac{N}{N/6} \leq 6. \quad (128)$$

Thus, in general, in this regime, we can upper bound the Gap by the constant 6.

• **Regime 2 ($1 < M \leq 0.1250N$) :**

For this regime, setting $s = \lfloor 0.5N/M \rfloor$, $\beta = 1$ in (116), we have

$$\begin{aligned} R_{\text{d2d}}^*(M) &\geq \frac{N \left(\frac{2\beta}{1+\beta} - s \frac{M}{N} \right)}{\frac{\beta N}{s} + 1} \times \frac{K}{K-s} \geq \frac{N \left(1 - \lfloor 0.5N/M \rfloor \frac{M}{N} \right)}{\frac{N}{\lfloor 0.5N/M \rfloor} + 1} \\ &\geq \frac{N(1-0.5)}{\frac{N}{0.5N/M-1} + 1} \geq \frac{0.5N(0.5N/M-1)}{N + \frac{0.5N}{M} - \frac{1}{N}} \geq \frac{0.5(0.5N/M-1)}{1 + \frac{0.5}{M}} \geq \frac{N}{M} \frac{0.5(0.5 - \frac{M}{N})}{1 + 0.5} \\ &\geq \frac{N}{M} \frac{0.5(0.5 - 0.1250)}{1 + 0.5} \geq \frac{N}{8M} \end{aligned} \quad (129)$$

Again, from [11, Theorem 1], we have

$$R_{\text{d2d}}(M) \leq \frac{N}{M} - 1 \leq \frac{N}{M}. \quad (130)$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq 8. \quad (131)$$

• **Regime 3 ($0.1250N < M \leq N$) :**

For this regime, setting $s = 1$ and $\beta = 1$ in (116), we have

$$R_{\text{d2d}}^*(M) \geq \frac{N-M}{N} = \left(1 - \frac{M}{N}\right) \quad (132)$$

Again, from [11, Theorem 1], we have

$$R_{\text{d2d}}(M) \leq \frac{N}{M} \left(1 - \frac{M}{N}\right) \leq \frac{1}{0.1250} \left(1 - \frac{M}{N}\right). \quad (133)$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq \frac{1}{0.1250} = 8. \quad (134)$$

Thus, for the case of $K > N$, the gap between the upper and lower bounds is 1 at $M = N/K$, at most 6 for $N/K \leq M \leq 1$ and at most 8 for all $1 < M \leq N$.

• **Case 3 ($N \geq K$)** : For this case, we consider two main sub-cases, namely (i) $K \leq 8$; and (ii) $K \geq 9$. We treat each case separately.

• **Sub-case 1 ($K \leq 8$)** : For this case, from [11, Theorem 1], we have

$$R_{\text{d2d}}(M) \leq \frac{N}{M} \left(1 - \frac{M}{N}\right) \leq K \left(1 - \frac{M}{N}\right), \quad (135)$$

where the last inequality is derived from the minimum storage constraint $KM > N$. Again, from (116), setting $s = 1, \beta = 1$, we have:

$$R_{\text{d2d}}^*(M) \geq \frac{N \left(1 - \frac{M}{N}\right)}{\lceil N \rceil} \geq \left(1 - \frac{M}{N}\right). \quad (136)$$

Thus the gap between the lower and upper bound is given by:

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq K \leq 8. \quad (137)$$

• **Sub-case 2 ($K \geq 8$)** :

For this case, we divide the cache memory into 3 distinct regimes namely, (i) Regime 1: $1 \leq N/K \leq M \leq 1.15N/K$; (ii) Regime 2: $1.15N/K < M \leq 0.1250N$; and (iii) Regime 3: $0.1250N < M \leq N$. We next consider each regime separately.

• **Regime 1 ($1 \leq N/K \leq M \leq 1.15N/K$)** :

For this regime, setting $s = \lfloor 0.4361K \rfloor$ and $\beta = 0.7398$ in (116) and using the fact that $K \geq 9$, we have

$$\begin{aligned} R_{\text{d2d}}^*(M) &\geq \frac{N \left(\frac{2\beta}{1+\beta} - s \frac{M}{N} \right)}{\frac{\beta N}{s} + 1} \times \frac{K}{K-s} \geq \frac{N \left(\frac{2 \times 0.7398}{1+0.7398} - \lfloor 0.4361K \rfloor \frac{M}{N} \right)}{\frac{0.7398N}{\lfloor 0.4361K \rfloor} + 1} \times \frac{1}{1 - \frac{\lfloor 0.4361K \rfloor}{K}} \\ &\geq \frac{N \left(\frac{2 \times 0.7398}{1+0.7398} - 0.4361 \frac{KM}{N} \right)}{\frac{0.7398N}{0.4361K-1} + 1} \times \frac{1}{1 - \frac{0.4361K-1}{K}} \\ &\geq \frac{N (0.4361K - 1) \left[\frac{2 \times 0.7398}{1+0.7398} - 0.4361 \frac{KM}{N} \right]}{0.7398N + 0.4361K - 1} \times \frac{1}{1 + \frac{1}{K} - 0.4361} \\ &\geq K \frac{\left(0.4361 - \frac{1}{9}\right) \left[\frac{2 \times 0.7398}{1+0.7398} - 0.4361 \frac{KM}{N} \right]}{0.7398 + 0.4361 \frac{K}{N} - \frac{1}{N}} \times \frac{1}{1 + \frac{1}{9} - 0.4361} \\ &\geq K \frac{\left(0.4361 - \frac{1}{9}\right) \left[\frac{2 \times 0.7398}{1+0.7398} - 0.4361 \times 1.15 \right]}{0.7398 + 0.4361} \times \frac{1}{1 + \frac{1}{9} - 0.4361} \geq \frac{K}{7}. \end{aligned} \quad (138)$$

Again, from (135), we have

$$R_{\text{d2d}}(M) \leq K. \quad (139)$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq \frac{K}{K/7} = 7. \quad (140)$$

• **Regime 2** ($1.15N/K < M \leq 0.1250N$) :

For this regime, setting $s = \lfloor 0.4470 \frac{N}{M} \rfloor$ and $\beta = 0.8995$ in (116), we have

$$\begin{aligned} R_{\text{d2d}}^*(M) &\geq \frac{N \left(\frac{2\beta}{1+\beta} - s \frac{M}{N} \right)}{\frac{\beta N}{s} + 1} \times \frac{K}{K-s} \geq \frac{N \left(\frac{2 \times 0.8995}{1+0.8995} - \lfloor 0.4470 \frac{N}{M} \rfloor \frac{M}{N} \right)}{\frac{0.8995N}{\lfloor 0.4470 \frac{N}{M} \rfloor} + 1} \geq \frac{N \left(\frac{2 \times 0.8995}{1+0.8995} - 0.4470 \right)}{\frac{0.8995N}{0.4470 \frac{N}{M} - 1} + 1} \\ &\geq \frac{(0.4470 \frac{N}{M} - 1) \left[\frac{2 \times 0.8995}{1+0.8995} - 0.4470 \right]}{0.8995 + \frac{0.4470}{M} - \frac{1}{N}} \geq \frac{N \left(0.4470 - \frac{M}{N} \right) \left[\frac{2 \times 0.8995}{1+0.8995} - 0.4470 \right]}{0.8995 + \frac{0.4470K}{1.15N}} \\ &\geq \frac{N (0.4470 - 0.1250) \left[\frac{2 \times 0.8995}{1+0.8995} - 0.4470 \right]}{0.8995 + \frac{0.4470}{1.15}} \geq \frac{N}{8M} \end{aligned} \quad (141)$$

Again, from [11, Theorem 1], we have

$$R_{\text{d2d}}(M) \leq \frac{N}{M} - 1 \leq \frac{N}{M}. \quad (142)$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq 8. \quad (143)$$

• **Regime 3** ($0.1250N < M \leq N$) :

For this regime, setting $s = 1$ and $\beta = 1$ in (116), we have

$$R_{\text{d2d}}^*(M) \geq \frac{N-M}{N} = \left(1 - \frac{M}{N} \right) \quad (144)$$

Again, from [11, Theorem 1], we have

$$R_{\text{d2d}}(M) \leq \frac{N}{M} \left(1 - \frac{M}{N} \right) \leq \frac{1}{0.1250} \left(1 - \frac{M}{N} \right). \quad (145)$$

Thus in this regime, the gap between the upper and lower bounds is given by

$$\text{Gap} = \frac{R_{\text{d2d}}(M)}{R_{\text{d2d}}^*(M)} \leq \frac{1}{0.1250} = 8. \quad (146)$$

Thus, for the case of $N \geq K$, the gap between the upper and lower bounds is at most 8 for all $N/K \leq M \leq N$. Combining the three cases completes the proof of Theorem 6. \square

REFERENCES

- [1] E. Bastug, M. Bennis, and M. Debbah, “Living on the edge: The role of proactive caching in 5G wireless networks,” *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, Aug 2014.
- [2] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [3] —, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” *IEEE/ACM Transactions on Networking*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015. [Online]. Available: <http://dx.doi.org/10.1109/TNET.2014.2317316>
- [4] U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands,” in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, April 2014, pp. 221–226.
- [5] R. Pedarsani, M. Maddah-Ali, and U. Niesen, “Online coded caching,” *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, April 2016.
- [6] A. Sengupta, R. Tandon, and T. C. Clancy, “Fundamental limits of caching with secure delivery,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 355–370, Feb 2015.

- [7] —, “Secure caching with non-uniform demands,” *IEEE Global Wireless Summit (GWS)*, pp. 1–5, May 2014.
- [8] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, “Learning distributed caching strategies in small cell networks,” in *Proc. The Eleventh International Symposium on Wireless Communication Systems (ISWCS)*, Aug 2014, pp. 917–921.
- [9] M. Ji, G. Caire, and A. F. Molisch, “Optimal throughput-outage trade-off in wireless one-hop caching networks,” in *Proc. IEEE International Symposium on Information Theory*, July 2013, pp. 1461–1465.
- [10] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, “Caching and coded multicasting: Multiple groupcast index coding,” in *In Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 881–885.
- [11] M. Ji, G. Caire, and A. F. Molisch, “Fundamental limits of caching in wireless D2D networks,” *arXiv:1405.5336*, 2014.
- [12] J. Hachem, N. Karamchandani, and S. Diggavi, “Multi-level coded caching,” in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2014, pp. 56–60.
- [13] J. Hachem, N. Karamchandani, and S. N. Diggavi, “Coded caching for heterogeneous wireless networks with multi-level access,” *arXiv: 1404.6560*, 2014. [Online]. Available: <http://arxiv.org/abs/1404.6560>
- [14] S. P. Shariatpanahi, A. S. Motahari, and B. H. Khalaj, “Multi-server coded caching,” *arXiv: 1503.00265*, 2015. [Online]. Available: <http://arxiv.org/abs/1503.00265>
- [15] Z. Chen, P. Fan, and K. B. Letaief, “Fundamental Limits of Caching: Improved Bounds For Small Buffer Users,” *arxiv:1407.1935*, August 2014.
- [16] M. M. Amiri and D. Gunduz, “Fundamental limits of caching: Improved delivery rate-cache capacity trade-off,” *arXiv:1604.03888*, 2016. [Online]. Available: <http://arxiv.org/pdf/1604.03888v1.pdf>
- [17] K. Wan, D. Tuninetti, and P. Piantanida, “On caching with more users than files,” *CoRR*, vol. abs/1601.06383, 2016. [Online]. Available: <http://arxiv.org/abs/1601.06383>
- [18] A. Sengupta, R. Tandon, and T. C. Clancy, “Decentralized caching with secure delivery,” *IEEE International Symposium on Information Theory (ISIT)*, pp. 41–45, July 2014.
- [19] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, “Caching and coded multicasting: Multiple groupcast index coding,” *arXiv*, vol. 1402.4572, 2014.
- [20] M. Ji, A. Tulino, J. Llorca, and G. Caire, “Caching-aided coded multicasting with multiple random requests,” in *In Proc. IEEE Information Theory Workshop (ITW)*, April 2015, pp. 1–5.
- [21] A. Sengupta and R. Tandon, “Beyond cut-set bounds-the approximate capacity of D2D networks,” in *Proc. Information Theory and Applications (ITA)*, February 2015, pp. 78–83.
- [22] A. Sengupta, R. Tandon, and T. Clancy, “Improved approximation of storage-rate tradeoff for caching via new outer bounds,” in *Proc. IEEE International Symposium on Information Theory*, June 2015, pp. 1691–1695.
- [23] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Hoboken, NJ, USA: Wiley-Interscience, John Wiley and Sons. Inc., 2006.
- [24] K. Wan, D. Tuninetti, and P. Piantanida, “On the optimality of uncoded cache placement,” *arXiv preprint arXiv:1511.02256*, 2015.
- [25] N. Ajaykrishnan, N. S. Prem, V. M. Prabhakaran, and R. Vaze, “Critical database size for effective caching,” *arXiv:1501.02549*, 2015.
- [26] H. Ghasemi and A. Ramamoorthy, “Improved lower bounds for coded caching,” *arXiv:1501.06003*, 2015.
- [27] C. Tian, “A note on the fundamental limits of coded caching,” *arXiv:1503.00010*, 2015.
- [28] C.-Y. Wang, S. H. Lim, and M. Gastpar, “A new converse bound for coded caching,” *arXiv:1601.05690*, 2016. [Online]. Available: <http://arxiv.org/abs/1601.05690>